



**DEVELOPMENT OF PRODUCTIVE CAPACITY
RELATIONSHIPS**

**Brice M. Stone
Kathryn L. Turner
Vincent L. Wiggins**

**Metrica, Incorporated
10010 San Pedro Avenue, Suite 400
San Antonio, TX 78216**

**Mary J. Skinner
Larry T. Looper
Jeffrey H. Grobman**

**HUMAN RESOURCES DIRECTORATE
MANPOWER AND PERSONNEL RESEARCH DIVISION
7909 Lindbergh Drive
Brooks AFB TX 78235-5352**

June 1996

Interim Technical Paper for Period February 1992 - December 1993

Approved for public release; distribution is unlimited.

**AIR FORCE MATERIEL COMMAND
BROOKS AIR FORCE BASE, TEXAS**

19961106 128

DEPT QUALITY INSPECTED 1

**ARMSTRONG
LABORATORY**

NOTICES

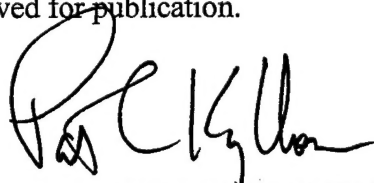
When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Office of Public Affairs has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

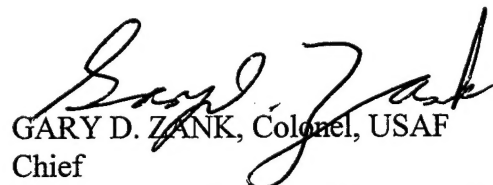
This paper has been reviewed and is approved for publication.



LARRY T. LOOPER
Project Scientist



PATRICK C. KYLLONEN
Technical Director
Manpower and Personnel Research Division



GARY D. ZANK, Colonel, USAF
Chief
Manpower and Personnel Research Division

Please notify this office, AL/HRPP, 7909 Lindbergh Drive, Brooks AFB TX 78235-5352, if your address changes, or if you no longer want to receive our technical reports. You may write or call the STINFO office at DSN 240-3853 or commercial (210) 536-3853.

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.				
1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE June 1996		3. REPORT TYPE AND DATES COVERED Interim Technical Paper - February 1992-December 1993
4. TITLE AND SUBTITLE Development of Productive Capacity Relationships			5. FUNDING NUMBERS C - F33615-91-D-0010 PE - 62205F PR - 7719 TA - 20 WU - 26	
6. AUTHOR(S) Brice M. Stone Mary J. Skinner Kathryn L. Turner Larry T. Looper Vincent L. Wiggins Jeffrey H. Grobman				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Metrica, Incorporated 8301 Broadway, Suite 215 San Antonio, TX 78209			8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Armstrong Laboratory (AFMC) Human Resources Directorate Manpower and Personnel Research Division 7909 Lindbergh Drive Brooks Air Force Base, Tx 78235-5352			10. SPONSORING/MONITORING AGENCY REPORT NUMBER AL/HR-TP-1996-0006	
11. SUPPLEMENTARY NOTES Armstrong Laboratory Technical Monitor: Larry T. Looper (210) 536-3648				
12a. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.			12b. DISTRIBUTION CODE	
13. ABSTRACT (Maximum 200 words) This research effort produced a job performance measure based upon data from Air Force occupational surveys. Core tasks are defined for each Air Force specialty (AFS) based upon the occupational survey data. The core tasks are used to determine the level of productivity for enlisted personnel in each AFS. The performance measure developed accounts for the number and difficulty of tasks performed by an airman. Productivity equations were estimated using this performance measure with aptitude and experience being the predictor variables. The performance measure was validated against a performance measure from a more traditional productivity data source. The performance measure was able to rank order airmen similar to the more traditional measure it was compared against. The core task methodology was then extended to 17 AFSs and performance measures and productivity equations were estimated for each AFS.				
14. SUBJECT TERMS Employee productivity Task difficulty Occupational analysis Performance measurement			15. NUMBER OF PAGES 68	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UL	

TABLE OF CONTENTS

SUMMARY	1
INTRODUCTION	1
LITERATURE REVIEW	2
Productivity/Performance Measurement	3
Theoretical Models	8
Empirical Models	10
METHODOLOGIES FOR MEASURING JOB PERFORMANCE	19
Core Task Concept	20
Methods for Determining Core Tasks	22
OMS Productivity Index	24
ESTIMATION AND VALIDATION OF PRODUCTIVITY	
INDEX USING EIGHT WTPT AFSs	29
Variable and Sample Definitions	29
Determination of Core Tasks	30
Estimation of OPI Equations	30
Coefficient Comparison Between E-Value and SME Method	36
Estimation of WTPT Proficiency Equation	38
Comparative Statistics (Correlations, Spearman Rank Order, Kendall's-tau)	41
Magnitudinal Effect of Experience and Aptitude	42
EXTENSION TO OTHER AFSs	45
CONCLUSIONS	53
RECOMMENDATIONS	54
REFERENCES	57

LIST OF TABLES

1. Conceptual Productivity/Performance Definitions	4
2. E-value Method Example	25
3. OPI Component Value Possibilities	27
4. Summary of Definition of Terms	29
5. AFS Titles and OMS Sample Sizes	30
6. Number of Core and Total Tasks	31
7. Descriptive Statistics for the OPI by Core Task Identification Method	32
8. Descriptive Statistics for TAFMS and Aptitude from OMS Sample	33
9. OPI Estimation Results -- E-value Method	34
10. OPI Estimation Results -- SME Method	35
11. Comparison of Standardized Coefficients for SME and E-value Methods	37
12. Descriptive Statistics for APF in the JPM Sample	38
13. Descriptive Statistics for TAFMS and Aptitude in the JPM Sample	39
14. APF Estimation Results	41
15. Comparison Statistics -- E-value Method	43
16. Comparison Statistics -- SME Method	43
17. Comparison of Standardized Coefficients for JPM and E-value Method	44
18. Definition of Clusters	46
19. Core Tasks	48
20. Descriptive Statistics for OPI in the Cluster AFSs	49
21. Descriptive Statistics for Experience and Aptitude in the Cluster AFSs	50
22. OPI Estimation Results for the Cluster AFSs	52
23. Standardized Coefficient Values for OPI	54

LIST OF FIGURES

1. Mean hands-on performance test (HOPT) scores by aptitude and job experience. Source: OASD (1989).	13
2. Comparison of Participation Rates and E-values	23
3. Example Productivity Function Using OPI	36

PREFACE

This research effort was conducted as delivery order number 0001 under Contract F33615-91-D-0010, by Metrica, Inc. for the Manpower and Personnel Research Division of the Armstrong Laboratory's Human Resources Directorate. The purpose of this effort was to develop productive capacity relationships for enlisted personnel.

The authors wish to express appreciation to Mr. Darryl Hand (computer programmer) and Mr. Kevin Borden (computer programmer) for their valuable technical contributions to this effort. The authors also wish to express appreciation to Dr. Walter Driskill and Mr. Johnny Weissmuller for their valuable technical contributions.

Major Sheree Engquist and 1Lt Jeff Grobman served as monitors of this contract effort for the Manpower, Personnel, and Training Integration Branch, Manpower and Personnel Research Division, Human Resources Directorate. Technical guidance and oversight was provided by Dr. William E. Alley. Dr. Mark Teachout provided data on Walk Through Performance Tests used in the validation phase and consulted on the proper use and interpretation of measures collected by the Air Force during the joint-service job performance project. 2Lt Janelle Viera and 2Lt Kevin Basik contributed to the technical review and editing of the final report.

SUMMARY

The objective of this research effort was to develop and quantify relationships between a measure of an airman's work productivity and his/her aptitude and experience in selected Air Force Specialties (AFSs). Occupational survey data were investigated as a potential source of a productivity index. Regression analysis was used to determine the relationship of the productivity indices with aptitude and experience factors. The performance measure developed in this effort was found to provide consistent rank orderings of enlisted personnel when compared to performance measures based on more traditional performance data. The research effort identified a methodology for determining core tasks for AFSs for which occupational survey data were available. These core tasks defined for each AFS were the basis for determining the productivity or proficiency of airmen within the AFSs. Airman performing all tasks defined as core tasks were considered to be fully productive. The productivity index developed from the core task concept accounted for the possibility of airmen performing tasks in addition to the core tasks. The performance measure developed showed statistically significant relationships with both aptitude and experience. The occupational survey data-based productivity measures were then compared with existing measures of productivity from the Job Performance Measurement Research Program with strong positive correlations. The measure was then successfully extended to 17 additional Air Force jobs covering the range of Air Force occupations.

INTRODUCTION

Measurement of the ability of an airman to produce valuable products and services, which contribute toward mission readiness and performance in combat, is central to any assessment of the force. Several theories of job performance or productive capacity have been postulated (Vroom, 1964; Porter & Lawler, 1968; Flamholtz, 1985). Some of these theories have been operationalized to determine optimal allocation of Air Force accessions based on aptitude. Carpenter, Monaco, O'Mara, & Teachout (1989) developed a selection method for a single Air Force specialty (AFS) which was based largely on a measure of productivity called "productive capacity." This methodology was extended by Faneuff, Valentine, Stone, Curry, & Hageman (1990) to include selection and allocation of recruits among multiple career fields. Others have used productive capacity in developing utility based valuations of force structures resulting from computer simulations of the airman inventory (Stone, Turner, Fast, Curry, Looper, & Engquist, 1991).

While each of these approaches relies heavily on estimates of an airman's productive capacity, these estimates were considered ancillary to the primary thrust of each research project. Productive capacity estimates were in support of the ongoing research.

Many of these productive capacity studies used job performance data from surveys as sources of data for model estimation. Many surveys of job performance are labor and time intensive and therefore quite costly to perform. This has limited the number of AFSs on which they could be performed, as well as the number of airmen participating in the studies. The

potential can be seen for a source of job performance data which is relatively low cost and available for all AFSs and large numbers of airmen. This job performance data must also be shown to be related to predictor variables such as aptitude and experience. The occupational data surveys used by the Air Force were investigated in this research effort as a potential source of job performance data upon which productivity measures could be established. Given the importance of productivity and productive capacity to the Air Force's ability to select, train, and use its people in the best way possible, a review of key productivity and occupational survey literature is necessary to establish a firm foundation for the use of occupational data to determine productive capacity.

LITERATURE REVIEW

The literature surrounding productivity is broad based and highly interdisciplinary with contributions from management science, psychology, and economics. In addition, there are many different facets to the analysis of productivity. The three aspects of productivity most relevant to the current project involve measurement, theoretical models of productivity determinants, and empirical models. Each of these will be reviewed below. However, given the varied terminology and goals evident in the literature, a brief conceptual overview will be useful.

Models and measurements for productivity from different disciplines (and often within disciplines) tend to have different terminologies and goals. In addition, it can be difficult to agree on what a productivity measure represents or how a particular conceptual definition of productivity should be measured. To provide a common context for the literature review, three measurable concepts of "productivity" will be used as touchstones:

- (1) productivity,
- (2) performance, and
- (3) proficiency.

Rarely will any single definition of these terms perfectly match the concept of productivity/performance used in a particular study. In addition, a large degree of conceptual overlap can be found in the definitions which follow. However, these conceptual definitions align with the most commonly used methods of measuring productivity and each term represents different aspects of productivity/performance. They will also provide a common nomenclature for classifying the theoretical and empirical models discussed.

Productivity typically represents some measure of output per unit of time. This output may be represented by a physical quantity (the number of tires replaced in an hour) or a more abstract or derived quantity (number of flights without a failure). In general, this concept of productivity is most directly applicable to production jobs with a measurable output and is somewhat more difficult to apply to many AFSs. As described here, productivity is generally aligned with direct measurements such as time and motion studies or piecemeal work. However,

as in the case of the number of flights without a failure, it can be employed by derivation from a measurable quantity.

Performance, in its common use (and as defined here), is usually a more general concept. It often includes overall contribution to organizational goals as well as some consideration of specific productivity as defined above. For example, performance may include interpersonal skills and the ability to improve the effectiveness of other workgroup members. It may also include some implicit consideration of fitness for promotion or the ability to cross-train into other areas. This definition is most applicable to supervisor or peer ratings. When applied to such ratings it may be preferable to consider two sub-categories of performance: organization performance and task performance. Organization performance applies to general ratings and those involving general skills. Conversely, the goal and operational definition of task ratings vary dramatically among studies; from a subjective measure of productivity to an often undefined performance concept. Even when the operational definition of a task rating uses the productivity definition above, it is likely that the responses will suffer from the rater's overall perception of the ratee's ability (i.e., the "halo" effect; Thorndike, 1920; Bass, Bernard, & Barrett, 1981). Given this effect, rated task performance will be considered under the general performance umbrella.

As defined here, proficiency will represent the demonstrated ability to perform a task or job (often within a specified time). In this sense, proficiency could be viewed as a somewhat weaker measure of productivity with time playing a smaller role. However, under this definition, proficiency is more amenable to jobs or specialties where the final output is difficult to quantify but the tasks or steps involved in correct performance can be observed. This measure is then representative of hands on testing where the objective is not maximum output; such as the Walk Through Performance Tests (Hedge & Teachout, 1986).

These three definitions are summarized in Table 1, and will be used to help compare some of the measurement methods as well as the theoretical and empirical models reviewed. Many of the theoretical models have some application to one or more of the three productivity/performance concepts. In other cases, one or more of these concepts are embedded within a model of abstract performance.

Productivity/Performance Measurement

Numerous methods of measuring productivity and/or performance have been proposed and applied. Almost every empirical study has employed a hand tailored method of measurement. Despite this variety, virtually all measurements are derived from one of two types of information: objective data or subjective data. While Landy and Trumbo (1980) include personnel data as a type of information, it is clear that most studies use personnel data as determinants of productivity rating rather than a performance criterion. In general, most objective data measurements share many characteristics and often suffer from the same potential deficiencies. Likewise, most subjective data measurements have a great deal in common.

Table 1. Conceptual Productivity/Performance Definitions

Productivity Term	Definition
Productivity	Output per unit time (either directly measurable or derived from other measures).
Performance	Contribution to organization goals (often subjective). Task performance is a subjective measure of ability to correctly (and possibly quickly) execute tasks.
Proficiency	Demonstrated ability to perform a task or job (perhaps within a prescribed time).

Objective Data Measurements

Objective performance/productivity assessment always involves some form of quantified testing or evaluation of work product. Quantified testing methods range from time and motion studies of specific operations to timed performance of an entire task. They may also take the form of product or task quality evaluations under prescribed time constraints or peak performance. More commonly, objective production data are used to measure output by "simply counting the results of work" (Guion, 1965). In cases where output is easily measured and all workers are operating under the same production conditions, this is an extremely viable measure of performance. It is a direct measure of the primary production goal for a position or specialty. Objective data are typically used to measure productivity or proficiency as defined earlier. Despite the benefit of being a direct and quantifiable measure of some aspect of job performance, objective data can have some limitations and inherent problems.

First, and perhaps most important, there are also many jobs for which objective measurements cannot completely capture an individual's productivity or performance (Bass et al., 1981). This is especially true of management positions or other positions for which it is difficult to quantify output (or even the tasks required for successful performance). In a related area, objective measures rarely capture the complete contribution of an individual to overall organizational goals. For example, objective measures seldom address a worker's ability to share expertise and thereby improve the productivity of co-workers. This argument suggests that subjective performance ratings may better measure an individual's overall contribution if the rating system includes dimensions for all organizationally important behaviors. However, as seen later, subjective measures suffer from their own set of potential difficulties.

Second, the results of objective measurement often include the impact of factors beyond the control of an individual (Barrett, 1966). Workers may have little control over the situation in which they are evaluated and these work conditions may have a direct impact on performance.

Great care must be taken to provide consistent conditions for testing or output evaluation. Lacking consistent conditions, the theoretical and empirical model must be able to account for differences in work conditions and these conditions must be measurable.

Subjective Data Measurements

Subjective productivity/performance measurement involves some form of rating by supervisors (although occasionally peer or self ratings are used). Subjective ratings typically cover one or more of three distinct areas: overall performance, dimensional performance, or task performance. Overall performance ratings seek to assess an individual's ability to complete all aspects of a position or their overall contribution to organization goals. Dimensional ratings focus the assessment of individuals on one or more dimensions or abilities which are crucial to job performance (e.g., interpersonal skills, technical knowledge). Task performance ratings are more closely aligned with objective measurements and are often used as a low cost substitute for costly objective testing. Supervisors rate the performance of job incumbents on specific tasks which comprise the incumbent's job description. Other than subjective task performance rating, subjective measures generally seek to capture an assessment of the ability to perform required behaviors rather than measure the results of behaviors as objective methods would. While subjective measurements are relatively easy to collect, they can be susceptible to various types of bias and inconsistency.

Eichel and Bender (1984) describe three overall subjective measurement methodologies which are applicable to a wide range of specific techniques:

- (1) Comparative methods which require the appraiser to evaluate the employees in a work unit relative to other workers in the unit.
- (2) Absolute methods which require the appraiser to evaluate the employee without making direct reference to the other employees.
- (3) Outcome oriented methods which require the appraiser to evaluate employees on the results they have achieved.

Comparative methods. Among the most common comparative methods are ranking, paired comparisons, and forced distribution systems. Each of these methods attempts to avoid the clustering of measurements associated with subjective methods. One of the most prominent ranking procedures is alternation ranking. Supervisors (or other raters) first identify their single best subordinate in the area being measured (e.g., a dimension or requirement of the job). Then, they identify their poorest subordinate. The second best and poorest workers are then identified. This process continues until all employees have been ranked and seeks to maximally discriminate employees performance along the specified measurement.

In the paired comparison method, raters are presented with a random pairing of their workers and are required to compare the one worker against each other worker. The number

of resulting pairs can be fairly large¹, and the method requires that the rater be familiar with the abilities of all of the individuals to be evaluated. An employee's rank is typically the total number of times he/she was chosen as the better of any pairing.

The forced distribution method requires that the measurements follow a specified distribution along a predetermined scale of effectiveness. The rater is forced to assign a certain percentage of workers to each of the defined categories. This implies that a discrimination among the workers must be made as if their performances conformed to the assumed distribution.

Absolute methods. Absolute methods do not require that a rater have a knowledge of the capabilities of all individuals to be assessed and in that sense impose less information requirements on the rater(s). However, this can significantly reduce the ability to use the measurements to compare performances among individuals, particularly if different raters evaluate different individuals. In general, the absolute methods are highly susceptible to most of the limitations of subjective measures described later. The most frequently used absolute methods include: weighted checklists, graphical rating scales, behavioral anchored rating scales, binary scoring, and mixed standard scales.

Weighted checklists give raters specific descriptive material to cover as they rate employees. As noted in Bass et al. (1981), one of the most commonly used methods is the Likert method of summing responses which allow raters to indicate whether they strongly agree, agree, are uncertain, disagree, or strongly disagree with a particular statement about an employee. Weights are assigned to each statement by raters or subject matter experts (SMEs) as to their importance relative to the jobs (Eichel & Bender, 1984).

Graphic rating scales are commonly used to assist the rater in making subjective measurements. Typically, several scales are developed representing various dimensions or aspects of job performance. Each of these dimensions is then placed on a rating scale. The scales vary dramatically in how many classifications (or steps) are provided for each performance dimension and in whether the steps are defined verbally, numerically, or not at all. These steps give the rater a simple choice for the performance level on which to rate the employees. The rater scores each performance characteristic or quality on a continuum from low to high for each employee. The method of deriving the dimensions of performance and the scales varies from simple fiat to the much more elaborate behavioral anchored system. Benjamin (1952), in a survey of over 100 merit rating plans, found that fewer than five steps will not allow raters to discriminate properly between employees.

Behavioral anchored rating scales (BARS) attempt to tie or anchor a rating system to specific behaviors relevant for job or task performance. In this manner, an attempt is made to control some of the subjective components in the measure. First, SMEs are asked to describe

¹The number of required comparisons can be calculated from the formula, $\text{comparisons} = n(n-1)/2$, where n = the number of individuals to be evaluated.

specific incidents critical to performance. These incidents are then clustered into five to ten performance dimensions related to general areas of performance. A second group of SMEs familiar with the job or position is asked to assign each incident to the dimension it fits best. An incident is typically retained in the design if 50-80% of both groups agree upon its dimension assignment. The result is a rating scale that has examples of behavior to anchor to each degree.

The method of critical incidents (Bass et al., 1981) involves aspects of both performance measurement and job/task analysis. SMEs identify situations in which they have seen a worker perform actions which are effective or ineffective and these incidents are then content-analyzed and categorized. These categorizations and their subcategories form the basis of a performance record to be used by the supervisor (or other rater). The supervisor records any incidents observed during a given period and at the end of a specified time block the records are analyzed for performance. The resulting data bank of incidents can be useful in providing information about job and organizational problems as well as performance.

Binary scoring (or forced choice checklist) methods are designed to correct for leniency and have typically demonstrated fairly high reliability (Wexley & Yukl, 1977). Four statements are constructed for each of 10 to 20 aspects of a job. Two of these four statements appear positive while two appear unfavorable. However, the statements are constructed such that only one of the positive statements actually discriminates effective from ineffective behavior (likewise for the two negative statements). The rater is required to choose two statements from each tetrad for an individual; the statement which best represents the individual's behavior and the statement which least represents the individual. Aggregation of responses makes it possible to discriminate among personnel along a performance continuum.

As described in Saal and Landy (1977), the mixed standard scale involves the assessment of employee attributes or traits. The scale presents the rater with three different degrees of the trait or attribute to be rated: a large amount of the trait, a moderate amount and finally one that states the trait is very low or absent. By presenting a large number of these descriptions to the rater randomly, it is possible to systematically identify those ratings that are inconsistent or do not follow a logical pattern.

Outcome oriented methods. Most of the outcome oriented measures discussed by Bass et al. (1981) are more closely aligned with objective measures as presented here. They include: direct indices related to appropriate measures of output or results; standards of performance relating accomplishments to a detailed set of expectation, and the more general management by objective.

Limitations and distortions of subjective measurement. As described in Bass et al. (1981) and Wexley and Yukl (1977) subjective measurements are subject to distortions from several sources. These include the clustering of ratings associated with leniency, strictness, and central tendency; as well as errors associated with the primacy effect, the halo effect, and rater characteristics. Leniency, strictness, and central tendency all represent the disposition of some

raters to cluster all responses at a particular level and fail to use the entire rating scale. This tendency makes it particularly difficult to compare workers evaluated by different raters.

The primacy and halo effects represent a tendency on the part of raters to utilize information other than that which directly bears on the performance in question. Primacy occurs when an individual who is initially successful in performing a certain task declines in performance but is still judged to have more potential than those individuals who start slow and then improve or fluctuate randomly. The memory of early performance dominates the ratings of current performance. The halo effect (Thorndike, 1920) occurs when a rater develops a global or overall perception about an individual which colors any judgments of specific attributes or capabilities. Rater characteristics such as the rater's ability to perform his/her job (Kirchner & Reisberg, 1962) or the rater's analytical ability (Korman, 1970) have been shown to have an effect on rater accuracy.

Theoretical Models

Theoretical models of productivity or performance provide a framework on which empirical models can be constructed. Theoretical models postulate determinants of performance as well as the relations among the determinants in producing performance levels. These descriptions can be tested to determine whether important factors have been left out or whether additional factors should be considered. They also offer testable a priori assumptions about the direction in which determinants can be expected to influence performance. More generally, they provide a constructive and criterion based framework against which empirical estimates can be compared for consistency. Taken collectively, the constructive and criterion based framework allows analysis of a model's epistemic validity (the relationship between actual criteria and underlying constructs). Likewise the framework allows examination of a model's constitutive validity (the postulated relationship between constructs). Most theoretical models of performance have been derived either from the chiefly psychological concepts of expectancy theory or economics based human capital theory.

Expectancy Theory

Several expectancy theory based models of job performance or productive capacity have been postulated and could serve as the basis for a model of productivity, performance, or proficiency estimates. Vroom (1964) and Porter and Lawler (1968) originally proposed expectancy theory with an emphasis on motivation, role perceptions, and ability. As formulated by Porter and Lawler (1968), ability is assumed to be based on stable individual characteristics such as personality traits, intelligence, and manual skills (Heneman & Schwab, 1972).

In a series of articles culminating in a book, Flamholtz (1985) outlined a methodology for human resource accounting which extended and refined the expectancy model of performance. Flamholtz views productivity or performance as one of three elements of an individual's

conditional value to an organization². The other two components of conditional value are assumed to be promotability (or the person's capability for promotion in the current work area) and transferability (or the capability for cross-training into other areas). Conditional value, and thus productivity, is postulated to depend on both individual and organizational factors. The individual factors considered are individual skills and activation level³ while the organizational factors are role and rewards.

Individual skills (such as technical, administrative, communication, etc.) are presumed to place upper limits to the nature and magnitude of services an individual can provide. As in Vroom's original model, these skills are assumed to be the products of cognitive ability and personality traits. Individual skills are assumed to be stable and enduring but can be modified by training. Activation level is associated with an individual's motivation and is not expected to remain constant. Both psychological and physiological traits are considered to be determinants of activation level. Operationally, activation serves to attenuate the expression of skills and prevent or enhance their full expression. Activation level is proposed to have a complex relation with both individual skills and organizational factors.

An individual's organizational role is hypothesized to be a third determinant of conditional value (and/or productivity). The role is a set of expected behaviors which provide an "opportunity to render potential services." Conditional value depends on the overlap of skills and role, but Flamholtz allows skills and activation to have an impact on roles. Conversely, activation is assumed to be a direct function of role through the vehicles of stimulus variation, intensity, complexity, uncertainty, and meaningfulness. Finally, an organization can affect productivity through rewards. Individual instrumental rewards are presumed to affect the worker's degree of activation, while instrumental system rewards increase the likelihood of staying in the organization. Most of these factors are assumed to have complex interactions with the other factors so that the factors are often co-dependent. Each factor, and the end result of conditional value, is dependent upon all individual and organizational factors.

Expectancy theory is a broad based model of performance/productivity. However, it does not consider the impact of factors of production such as capital availability (computers for office workers, tools for technicians, etc.) and only considers training in an ancillary fashion. While expectancy theory offers a wealth of possible determinants of performance, it also allows for an extremely wide latitude of interactions among the determinants with little guidance for empirical modelling.

²The value is only conditional in the sense that the individual must remain with the organization for the value to be expressed. Restated, the value is conditional on continuation with the organization.

³Flamholtz defines activation level as the "neuropsychological counterpart of the notion of motivation." In this sense, it includes not only psychological motivation, but also the metabolic and neurochemical context in which "motivation" is expressed.

Human Capital Theory

Human capital theory, as developed by Becker (1964, 1971), can be viewed as a direct extension of economic production theory (Becker, 1971) into the areas of human skills, training, and performance. The theoretical underpinnings of human capital theory imply that innate skills, education, training, job experience, and task experience are all factors in determining a person's human capital. This level of human capital then has a direct impact on a person's capability to perform a particular job and thus affects his/her performance or productivity. Further, this human capital can be used as a direct input into production functions in place of a simple accounting for labor hours or dollars.

The human capital model provides the framework for analyzing the role of individual productivity determinants (human capital) within the larger framework of production (which includes other factors of production such as facilities). This allows the analysis of the impact of human capital on organizational production or performance as well as the impact of facilities and other production factors on the effectiveness of existing human capital. By casting human capital into the well researched arena of production theory, a large body of research on production theory can be used to provide insight into specific model formulations. Production is a heavily researched area where many specific mathematical models have been developed, analyzed for their implications, and empirically estimated (e.g., Hicks, 1932; Arrow, Chenery, Minhas, & Solow, 1961; Intriligator, 1978). In addition to productivity, the theory has implications for optimal amounts of education and training (both from the organization and individual viewpoint; Becker, 1965), the transferability of training, labor force participation, and occupation choice.

Empirical Models

Many empirical models of task or occupational performance, productivity, and/or proficiency have been developed and estimated. Empirical models seek to quantify the relationship of performance to other factors or to measure interrelationship among several criteria or determinants. These models have varied from the ad hoc estimation of relations based on available data to models derived in varying degrees from expectancy theory or human capital theory.

Background Work

In some Air Force research, Wiley (1976) attempted to determine if an aggregation of task level subjective supervisor ratings was "superior" to overall supervisor ratings. Four sets of subjective measures were obtained for first term airmen in the 645X0 (Inventory Management) and 647X0 (Material Facilities) AFSS: ratings of (1) global performance, (2) 65 behavioral traits, (3) performance on tasks involved in the specialty, and (4) required training time for these same tasks. For each of 244 airmen, supervisor ratings were independently collected from two supervisors. Tests were performed on inter-rater similarity among the measures as well as the development of constructive models using grade, skill level, job difficulty index, task experience,

and task difficulty index. Overall, Wiley found that the performance of aggregates from the task inventory ratings had only slightly better inter-rater performance and constructive validity. From the data available, it was concluded that the slight improvement from job inventory ratings was not sufficient to overcome the additional cost and time requirements to obtain such detailed information.

Wiley and Hahn (1977), in another Air Force study, attempted to model overall supervisor task performance ratings in six areas (general performance, quantity, quality, exceeds expected work share, self-initiative, and shares knowledge) from an inventory of task ratings by the incumbent (self), peers, and supervisors. In addition, the model included data on incumbent grade, total active federal military service (TAFMS), selector aptitude indicators (AI) scores from the Armed Services Vocational Aptitude Battery (ASVAB), gender, and other demographic factors. In addition, scores from an experimental test battery of 11 short cognitive tests were included as determinants of the six overall performance ratings. Regression analyses relating all of the factors to each of the six overall performance ratings were performed for 3 AFSs⁴. While different determinants were found to be important for different specialties and across general performance areas within specialties, the researchers found several consistent patterns in the analyses (quoting the most relevant conclusions from the authors):

- (1) Raters agreed on task performance evaluations to a statistically significant degree⁵.
- (2) Raters agreed better on rating overall job performance than on rating the performance of single tasks.
- (3) Incumbents were poorer sources of task ratings than peers or supervisors.
- (4) Peers could be substituted for supervisors as performance evaluators without great loss in reliability.
- (5) A few task ratings taken together accounted for a substantial percentage of the overall performance rating variance.
- (6) Aptitude data and demographic data (such as grade and length of service) accounted for much less of the overall evaluation of an airman's performance than can be accounted for using ratings on as few as five tasks.
- (7) Some of the overall performance variance was attributable to the attitudes and interest of the incumbents.
- (8) Marked differences distinguished the 304X4 performance ratings from the other two specialties; they were internally more reliable and better able to predict overall performance ratings.

⁴The specialties examined were: 291X0, Telecommunications Operations Specialist; 304X4, Ground Radio Communications Equipment Repairman; and 431X1C, Operations Specialist. These were broken down into inventories of 51, 95, and 55 tasks respectively for the task level ratings.

⁵There is evidence for inter-rater reliability.

- (9) Use of the top of the rating scale was frequent when performance ratings were made on easier tasks, and on performance in AFSs with lower aptitude requirements.
- (10) Since measurability was better for more difficult tasks, with less use of the rating scale, the AFS with high aptitude incumbents received the lowest mean performance ratings.

The Wiley and Hahn model has a rather ad hoc specification and draws data from basically all available sources. In addition, the model and conclusions 5 and 6, seem to confuse the use of criterion variables (overall performance ratings and task performance ratings) and constructive determinants (length of service, aptitude, etc). Both theoretical models discussed earlier would consider the relation between overall performance, peer ratings, and task performance as a form of criterion validity as opposed to a constructive model explaining any of the three performance measures. Despite these drawbacks, the findings of Wiley and Hahn are often echoed in the results using more sophisticated models and methods.

Hands on Performance Testing

Congressional initiation in 1980 of a joint service effort to relate entrance standards to job performance sparked a flurry of military research on empirical performance/proficiency models. While there are differences in approach across the services, most of the current research centers around objective hands-on or walk through performance testing (WTPT). Each service has implemented variations of the hands on WTPT procedure but they all center on the evaluation of individual performance in completing critical tasks for a particular specialty or job. Performance is evaluated by a trained rater during hands on testing of the critical tasks. The task inventories, their relative importance, and objective methods of measuring performance are carefully developed using SMEs in the specific specialties. In addition to the hands on testing, most of the services are simultaneously collecting additional criterion measures such as subjective BARS ratings on overall or dimensional performance and technical knowledge tests. All of the initial efforts in the initiative have focused on first-term enlisted personnel. An overview of the methodologies for testing can be found in the annual reports of the Office of the Assistant Secretary of Defense (OASD, 1982-1989).

In the eighth annual report (OASD, 1989), descriptive summaries are presented of a cross-sectional analysis of all hands on performance data collected to that point. The data encompasses tests on over 7,700 enlisted personnel where the test scores have been standardized for each specialty in each service to improve comparability of the criterion measures. As seen in Figure 1, a compelling difference can be seen in the mean performance of personnel by aptitude. Significant hands-on performance differences are observed between mental category I-II, IIIa, IIIb, and IV, as measured by the Armed Forces Qualifying Test (AFQT) score of the ASVAB. Further, the experience-performance paths differ by mental category. Performance improves with experience across all mental categories, but the improvements are much more pronounced for personnel in lower mental categories. Consequently, the performance of low and high mental category individuals becomes more similar as experience increases.

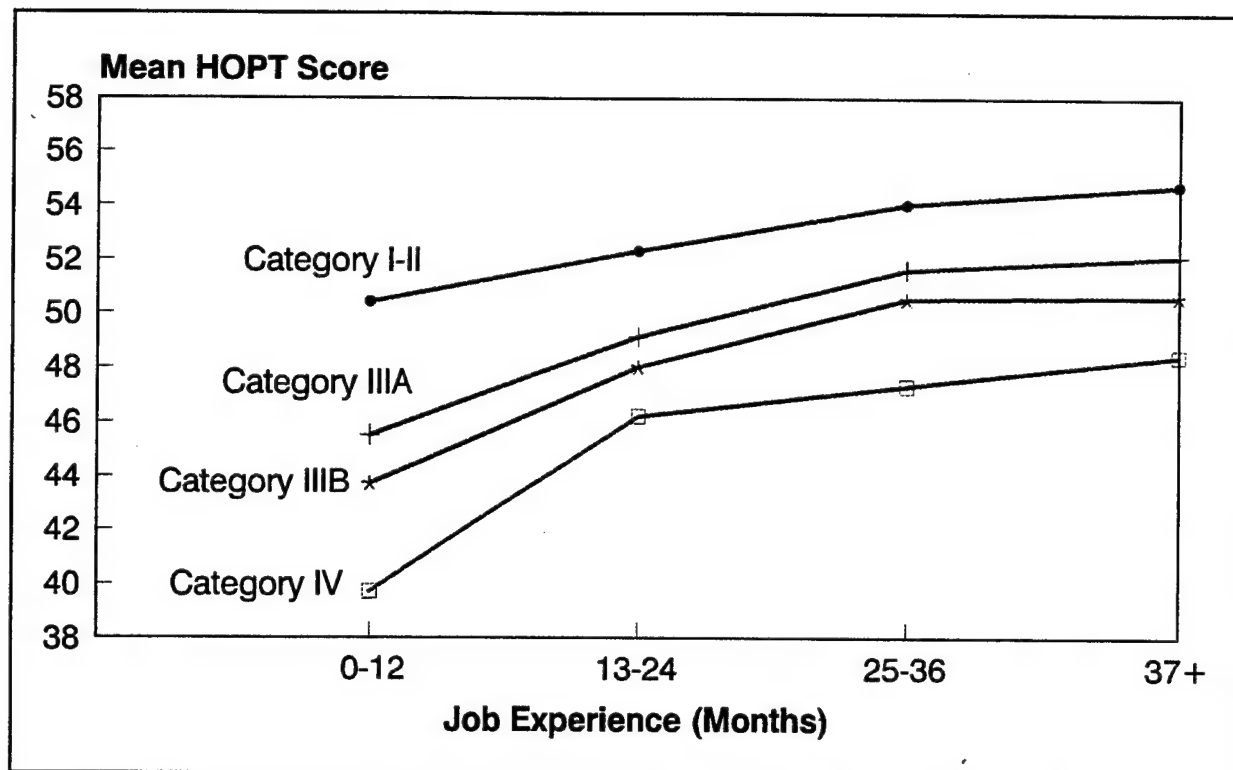


Figure 1. Mean hands-on performance test (HOPT) scores by aptitude and job experience.
Source: OASD (1989).

The Army collected hands-on test performance data on 9 military occupational specialties (MOS), and with subjective BARS and service wide performance ratings on an additional 10 MOSs (OASD, 1988). Unlike the other services, the Army derived core tasks which were intended to be applicable across specialties and were developed to reflect what were considered important dimensional components of performance (e.g., general soldiering composite, technical performance, personal discipline, etc.). The components have low inter-correlations and thus seem to be capturing independent performance dimensions. Using correlation analysis on the AFQT percentile score, aptitude was predictive of some dimensional aggregates of the hands on testing but not predictive of others (ARI, 1986). The aggregates that were dependent on AFQT were collectively labeled "can do" (technical performance and general soldiering); those that were not dependent on AFQT were labeled "will do" (effort/leadership, personal discipline, and fitness).

The Navy has reported results on two ratings (Machinists Mate and Radioman) for hands-on performance testing; subjective BARS ratings from peers, self, and supervisors; and knowledge tests (OASD, 1988, 1989). The hands on tests utilized 14 tasks for Machinists Mate and 12 tasks for Radioman. Unlike the Army, the Navy found a high degree of inter-relation among the tasks in a rating. Using correlation analysis adjusted for sample selection bias or

truncation⁶, the Navy found AFQT percentile and experience (time in service) each had a moderate impact on hands on performance criteria. However the BARS scores had a weaker relation to AFQT percentile (showing a weaker construct validity). Overall there were significant relationships among the hands on, BARS, and knowledge test criteria (indicating some criterion validity for all three measures). Equipment differences and differential access to on the job training (OJT) also seemed to play a role in performance levels. This is an implicit recognition of several factors considered to be performance determinants in human capital theory.

The Air Force has collected hands on WTPT data on eight specialties and also collected knowledge information, global performance ratings, dimensional performance ratings, Air Force wide performance ratings, and subjective task performance ratings (Hedge & Teachout, 1986; Lipscomb & Hedge, 1988). The annual OASD report published in 1988 emphasized correlation analysis on four AFSs with an adjustment for sample truncations. Experience was consistently found to be an important factor in hands on performance while the effect of AFQT was always significant but varied by specialty (education had little effect). An interesting result of the analysis of generalizability was that test subjects were rank ordered similarly across the four types of subjective ratings.

In an analysis of four infantry MOSs, the Marines found that AFQT percentile was highly correlated with hands on performance. Adjusting for sample truncation, they also found that experience was an important factor in hands on performance. In addition, the experience profiles for high aptitude (mental category I-IIIa) and low aptitude (IIIb-IV) infantrymen were different.

OASD has undertaken a much more involved study to attempt to operationalize the results of the hands on performance test. A nested model methodology has been developed for estimating the relation between hands on performance, aptitude, and experience, while simultaneously generalizing the results to untested specialties (Harris, McCloy, Dempsey, Roth, Sackett, Hedges, Smith, & Hogan, 1991). The process involves mapping 24 military occupations (9 Army, 3 Navy, 8 Air Force, and 4 Marine) to civilian jobs and using the Department of Transportation (DOT) set of 44 job characteristics to portray each of the 24 occupations. These 44 characteristics were factor analyzed and four common factors were extracted. Using this approach, the model could be applied to other specialties simply by creating the four common factor scores from a match to a DOT civilian job. These factor scores were then used as one set of determinants in a hierarchical model relating hands on performance to AFQT percentile, an ASVAB technical composite score, education level, and time in service.

Essentially the coefficients relating AFQT, technical composite score⁷, education, and time-in-service (TIS) to hands on performance become functions of the factor scores of the job characteristics. Thus coefficients can be generated for any military specialty which has been

⁶For any service, the tests were only performed on job incumbents who had already passed the required selection criteria to be assigned to the specialty.

⁷Like the AFQT, the technical composite score is derived from ASVAB subtest scores.

matched to a civilian job. Harris et al. (1991) went to great effort to improve the comparability of the hands on test scores among the different specialties and services so that all of the specialties could be used in a single estimating equation. Using a linear specification with allowances for some interactions among the factors, all of the input factors had significant direct impacts on hands on performance. The relationship between technical composite score and TIS was negative and statistically significant confirming the overall OASD results that the performance of high and low aptitude individuals grows closer as experience is gained. Finally, the researchers found that the effects of AFQT, technical composite, and TIS varied among many of the 24 specialties for which criterion data were available.

The studies discussed above all used specialty level hands on performance measures aggregated from the results on task level hands on testing. The number of tasks for a specialty varied from 10 to 95. Lance, Hedge, and Alley (1987) examined the Air Force's hands on WTPT data and several other criterion measures collected with the WTPT at the task level. Focusing on the Jet Engine Mechanic specialty (AFS 426X2); the Lance et al. study developed hierarchically moderated regression models (Arnold, 1982, 1984) relating criterion variables to aptitude (Mechanical (M) selector AI), task difficulty, and either a job experience composite (from TAFMS, months on engine, months in unit, shop experience, and flightline experience) or a task experience composite (from subjective task experience ratings and estimated times the task had been performed). The criterion variables considered were the objective WTPT score and subjective overall performance rating (by the WTPT rater), self task ratings, and supervisor task ratings. In regressions on 255 airman involving 15 tasks; Lance et al. found that virtually all of the determinants had significant main effects on each of the four criterion variables (however, a significant relation could not be found between overall performance ratings and task difficulty or job experience). An interaction between aptitude and experience indicated that the performance gap between high and low aptitude airmen narrowed with experience, supporting the OASD and Harris et al. (1991) results. In addition, Lance et al. (1987) found that the alternate criteria were relatively uncorrelated but had similar patterns of regression coefficients. This could be taken as an indication of construct validity and the researchers concluded that the alternate measures were capturing different aspects of the criterion space.

Alley and Teachout (1990) also found an empirical relationship between experience (measured by TAFMS), aptitude (measured by the best linearly weighted ASVAB subtest composite), and WTPT performance. Separate regression estimates were developed using aggregate data from each of the eight specialties involved in the Air Force's initial performance measurement effort. Aptitude was found to have a significant effect on performance in five of the eight specialties while experience was significant in seven of eight. Alley and Teachout (1990) tested for an aptitude and experience interaction, but found them significant for only one specialty. Schmidt, Hunter, and Outerbridge (1986) have also empirically confirmed the importance of both experience and aptitude on performance. These relationships were detected for both objective job knowledge criteria and supervisory ratings. Further, Schmidt, Hunter, Outerbridge, and Goff (1988) also found that the impact of aptitude was constant for those with low and high experience; that is, there was no interaction between aptitude and experience.

Using task level data for 684 airmen in four specialties from the same Air Force WTPT data, Ford, Sego, and Teachout (1991) came to substantially different conclusions. Using self reported task experience as a determinant, they found that aptitude⁸ (AFQT) and job experience were no longer predictive of WTPT performance. Using hierarchical regression, on each of 43 tasks, the researchers found that the coefficient on AFQT was significant in only 7 tasks. This result is in stark contrast to all of the other studies conducted among the services.

The Air Force WTPT data has also been used to derive empirical performance/proficiency models as components of larger systems. Faneuff et al. (1990)⁹ developed models of airman time to proficiency (TTP). Using WTPT data from six specialties, regression models were developed using airman skill level, experience (the log of TAFMS), and aptitude (either the appropriate selector AI percentile or AFQT percentile). Confirming the results of Alley and Teachout (1990) and other researchers, experience was a significant determinant of WTPT performance in four of six specialties. Likewise, selector AI percentile was significant in four of six specialties (when used in place of selector AI, AFQT was significant in 3 equations). Stone et al. (1991) extrapolated these results to all AFSs using clusters of AFSs.

Using the same six specialties and hold-out samples for cross validation, Wiggins, Looper, and Engquist (1991) found some evidence for more than linear structure in the experience, aptitude, WTPT performance relationship. Out-of-sample performance of neural network models as compared with regression and nonlinear logistic regressions provided weak evidence of unmodeled nonlinear or interacting structure in some of the AFSs. This is consistent with the often conflicting results from prior studies when testing for interactions between experience and aptitude.

Vance, MacCallum, Coover, and Hedge (1988) used three tasks from the Air Force WTPT data for Jet Engine Mechanics to examine the construct validity of the different criteria. Using confirmatory factor analysis (Widaman, 1985), they found significant convergence among the three sources of subjective performance rating (self, peer, and supervisor). This implies that self, peers, and supervisors are all valid sources of performance ratings. Additional criterion validity was established by the convergence between the subjective ratings and objective WTPT scores.

The validity work was extended to epistemic and constitutive validity in Vance, MacCallum, Coover, and Hedge (1989). Using covariance structure models, or nomological networks (Cronbach & Meehl, 1955), linear covariance structures were examined among the different criteria (WTPT score and self, peer, and supervisor task ratings) and constitutive model components (months on the job, times performed task, selector AI composites, and reported supervisor support). Strong evidence was found for both epistemic validity (the relation of the

⁸Referred to as "cognitive ability" in the paper.

⁹For more detail on the WTPT models see Stone, 1989.

measures to an underlying criterion) and constitutive validity (the relation between the underlying criterion and its hypothesized determinants). Vance et al. (1989) found that for three Jet Engine Mechanic tasks, experience was consistently important in explaining performance. However, selector AI aptitude was found to affect technical school performance but not the underlying criterion of job performance.

Horowitz and Sherman (1980) employed a human capital approach in deriving a model of performance for Navy personnel. They analyzed performance at the ship level using ship subsystem downtime as an inverse measure of production for the personnel. Analyses were conducted separately for five Navy ratings (Boiler technician, Machinists's mate, Fire-control technician, Gunners mate, Sonar technician, and Torpedoman's mate) and six subsystems on which the personnel worked (boilers, engines, etc). In general, any person only worked on one or two subsystems. Separate regressions were estimated for each of the sub-systems using the amount of equipment downtime as the criterion variable. Using a human capital approach, the condition of the equipment, complexity of the equipment, number of workers in the relevant rating(s), and characteristics of the seamen were considered as determinants of downtime per month. This differs from the research reviewed above in that other factors of production are considered directly in the model. Worker characteristics represented the personnel side of the production function and their effects were the primary target of the research. Personnel characteristics (such as education, entry test scores, etc.) were averaged over the personnel in the relevant rating(s) for the particular type of equipment. Horowitz and Sherman (1980) found several interesting relations: 1) length of service or amount of sea duty (experience) was important for all ratings; 2) education level was most important for ratings in highly technical areas; 3) entry test scores were important for three of the six ratings; and, 4) individuals in high pay grades contributed more. Perhaps most important, in most ratings high education levels, test scores, and training increased productivity most for those workers or seamen handling relatively complex equipment. In other words, task complexity was a factor in the relative importance of aptitude and training.

Use of Occupational Survey Data

It is clearly seen from these studies that knowledge about the details of what an airman does in a job (i.e., the tasks performed) and the relationship of the performance level of those tasks to the airman's aptitude and job experience is critical to a fuller understanding of worker productivity. The Air Force routinely collects, through mail surveys, an extensive set of data on the tasks airmen perform in each occupation. This survey of jobs is called the Occupational Measurement Survey (OMS). It is conducted by the Air Force Occupational Measurement Squadron every 3-7 years (or sooner if needed) on each AFS to determine training needs and help structure jobs. It holds potential for a generalizable model of enlisted force productivity.

Nathan and Nathan (1991) have produced some constitutive validity results which may have a significant bearing on the use of OMS data in performance measurement. They argue that most subjective performance measures implicitly assume that persons in the same jobs are performing the same tasks. Conversely, they contend that managers are unlikely to assign tasks

to workers who cannot adequately perform the task. Thus, an airman may be subjectively rated as performing his/her job well primarily because the tasks he/she does have been tailored to his/her own abilities. Arguing along these lines, they suggest that the number of tasks (i.e., different types of tasks) performed may be a more valid gauge of performance. This measure assumes that more capable individuals will be given a larger variety of tasks (it also implicitly assumes that the "easy" tasks are still counted in the capable individuals task inventory). This hypothesis was tested using three potential measurement criteria as supplied by supervisors:

- (1) overall performance,
- (2) inventory of tasks performed, and
- (3) inventory of tasks performed well.

Using a sample of 130 observations of clerical workers and four aptitude factors (three taken from nine short clerical tests and the fourth from a general mental ability test), regressions were estimated to model each of the three criteria. Using regression fit as a measure of constitutive validity, the inventory of tasks performed had the highest validity, the simple inventory had second highest, and the overall performance rating had the lowest overall validity. While the model may have lacked factors considered to be theoretically important and found important in other studies, principally experience, the results indicate that OMS data may have a role to play in performance/productivity/proficiency modeling.

Miller, Skinner, and Harville (1992) also investigated the possibility of using Air Force occupational surveys as a source for job performance indices. The researchers examined the relationship between the number of tasks performed by airmen and measures of job difficulty and aptitude. Three performance measures were computed including: number of tasks performed (NTP), Average Task Difficulty Per Unit Time Spent (ATDPUTS), and Job Difficulty Index (JDI). Predictor variables used included ASVAB select AI scores and number of months of job experience (obtained from the job inventory background information provided by survey respondents). The researchers found that experience effects were consistently seen across all specialties for the performance measures. The results indicated that the NTP and difficulty (ATDPUTS and JDI) of the tasks performed by airmen tended to increase for personnel with more experience. Aptitude effects were not as strong as those seen for experience, although significant positive relationships were seen for at least one performance measure in over half of the AFSSs investigated. The findings with experience and aptitude were consistent with those of Nathan (1992), and suggest that the utility of occupational surveys in assessing job performance should be further explored.

The findings of Nathan and Nathan (1991), Nathan (1992), and Miller et al. (1992) suggest that occupational survey data are a key source of job performance data. This present research effort will investigate the viability of using occupational survey data for developing measures of job performance. Aptitude and experience will then be explored as predictors of job performance developed from the occupational surveys.

METHODOLOGIES FOR MEASURING JOB PERFORMANCE

Several sources were investigated during the literature and data review phase of the project to determine their utility for the development of a measure of job performance. The two primary data sources decided upon were Job Performance Measurement (JPM) and OMS data. JPM data have been used as sources for job performance measures with promising relationships with experience and aptitude seen consistently in the literature. JPM data are a logical beginning point for the data investigation given the number of job performance measures which have been used as the basis of measures of productivity in previous research efforts (Stone et al., 1991; Faneuff et al., 1990; Carpenter et al., 1989; Harris et al., 1991). The JPM data, however, have several limitations for generalizing from the original eight AFSs to other AFSs in the Air Force:

- (1) The JPM data were collected on only eight 5-digit AFSs, two from each of the four selector AIs,
- (2) The JPM data have limitations when an estimated relationship for each or all of the eight AFSs is generalized to other AFSs. Alternative AFSs may not be sufficiently similar in duties and/or tasks to identify one of the eight AFSs as a proxy,
- (3) The JPM samples used for each AFS were small, limiting the type of analysis which could be performed to produce the productivity relationships (Wiggins et al., 1991),
- (4) Productivity relationships estimated with the JPM data have produced relatively poor statistical results in some studies (Carpenter et al., 1989; Faneuff et al., 1990), and
- (5) The JPM samples included only first-term airmen.

No other survey or personnel data has as many alternative measures of job performance as the JPM data. Techniques to extend the limited JPM data base to other AFSs have displayed only modest statistical significance (Stone et al., 1991; Faneuff et al., 1990; Carpenter et al., 1989; Harris et al., 1991). The cost of test development and data collection for the JPM data prohibit extending study to the other remaining enlisted AFSs.

OMS data could potentially offer a lower cost alternative to the JPM study data if job performance measures can be found within the occupational survey data. Initial success was seen in developing job performance measures and relating the measure to aptitude and experience in the Nathan and Nathan (1991), Nathan (1992), and Miller et al. (1992) studies. OMS data consists of job inventories for each AFS. The job inventory for each AFS is comprised of task statements which cover all tasks performed by airmen in that AFS. Data are collected on airmen responses to whether he/she performs or does not perform each task within the job inventories. Airmen also provide information related to the percent of time spent performing each task. Supervisors and SMEs rate the difficulty of performing each task within the inventory, as well as the training emphasis for each task. The OMS data offer several benefits and liabilities:

- (1) The OMS data are collected on all 5-digit AFSs, as well as all tasks, and updated periodically as the structure of the career field changes,
- (2) The OMS data are collected on airman possessing lengths of service from 1 day to 20 years or more, grades E2 through E9, skill levels 3, 5, 7, and 9,
- (3) The OMS data which are collected on each career field come from relatively large sample sizes, allowing more flexibility concerning the type of analysis which can be performed to produce the productivity relationships,
- (4) The OMS data have been used on a limited basis for the estimation of productivity relationships (Stone, Rettenmaier, Saving, & Looper, 1989; Stone, Turner, Engquist, & Looper, 1992), and
- (5) The OMS data offer few options as direct measures of productivity or proficiency.

In order to use the OMS data for the purpose of estimating productivity relationships, a methodology must be defined which would use the data elements to calculate a measure of productivity. This methodology must be generalizable to all 250+, 5-digit AFSs and be defensible as a measure of productivity. The OMS productivity measure developed as a result of this present research effort will be validated using the JPM data given the amount of previous research using the JPM data to estimate productivity relationships.

Core Task Concept

The analysis which follows was based upon a methodology which used the concept referred to as core tasks (Phalen & Weissmuller, 1991). Core tasks are those tasks which are done by a high proportion of the members of an AFS (which will be subsequently termed a high participation rate). For the purpose of developing a productivity measure, the assumption is that core tasks comprise a group of tasks associated with the normal performance of the job. An individual is expected to perform these core tasks to be fully productive in the career field. Individuals not performing these core tasks are assumed to be less than fully productive. New airmen initially entering a career field may or may not perform all the core tasks, but, as they gain job experience, they are expected to perform an increasing number of the core tasks as a part of their job responsibilities.

Thus, an individual who performs only a subset of the core tasks is considered to be less than fully productive. An individual who performs¹⁰ all the core tasks is considered to be fully productive; however, they are then expected to begin assuming additional non-core tasks. Thus, an individual performing all the core tasks and additional non-core tasks would be more than fully productive at performing the normal responsibilities of the job. This concept is consistent with the results seen in the Miller et al. (1992) study where number of tasks and difficulty of tasks tended to increase with the experience of airmen.

¹⁰Performance for all tasks, whether core or non-core is assumed to be at a satisfactory level.

The definition of less than, greater than, or fully productive based on the performance of core and non-core tasks is straight forward. In reality, the performance of tasks in the work place is not this orderly. A problem arises if individuals are not performing all the core tasks but are performing some non-core tasks. These individuals could be considered more than fully productive or less than fully productive depending on the rationale applied to the combination of (less than all) core tasks and non-core tasks exhibited by workers in the work place. This (less than all) core tasks and non-core tasks combination is a common occurrence as indicated by the task participation rates in the OMS data.

Task Difficulty as an Addendum to Core Tasks

To determine the productivity level of those individuals who exhibit a (less than all) core tasks and non-core tasks combination, a second factor, task difficulty, was considered as an addendum to the definition of the core tasks as a measure of productivity. Task difficulty is an index which indicates how difficult a task is to learn based on a 7-point scale. Assume that the total task difficulty associated with the core tasks defines a baseline, **B**, which is equal to

$$B = \sum_{i=1}^{n_c} TD_{i,c} \quad (1)$$

where

$TD_{i,c}$ is the task difficulty associated with i th task of the core tasks and
 n_c is number of core tasks.

An airman performing a (less than all) core tasks and non-core tasks combination can be classified as less than, greater than, or fully productive based on the total task difficulty associated with the tasks being performed compared to the baseline total task difficulty, **B**. An airman which exhibits a total task difficulty above the baseline level established by the core tasks would be considered more than fully productive, while an airman exhibiting a total task difficulty below the baseline level would be considered less than fully productive. In addition, an airman can compensate for not performing the core tasks, as well as for performing a fewer number of total tasks, by substituting more difficult tasks for the core tasks which are not being performed.

Identification of Core Tasks

The identification of the core tasks was completed using OMS data. OMS data provide information concerning which of the tasks associated with a particular career field has been performed by each respondent. The level of participation which determines which tasks are core varies among career fields. The identification of the core tasks is done by career field, category of enlistment, years of service, skill, and grade. Thus, core tasks are identified based on a stratified sample for each career field and the participation rates associated with the tasks.

Several caveats are apparent and supportable by the OMS data but not generalizable to all career fields:

- (1) Core tasks could differ by base location due to differences in duties and responsibilities, training procedures, or equipment from base to base,
- (2) Some career fields could display too much diversity in the performance of the job even within a particular location, e.g., 732x0 (Personnel), to concisely identify core tasks, and
- (3) Core tasks do not directly measure the proficiency with which a task is performed, only that the task is performed by the airman. Proficiency is captured through the performance of non-core tasks in addition to core tasks.

JPM data explicitly measure whether a task is performed correctly. Each task in the JPM data was split into individual steps included in the performance of the task. These steps were scored as correctly or incorrectly performed by each subject. The total walk-through-performance test (twtp) score was developed from these step-by-step scores (Hedge & Teachout, 1986) and is a measure of the proficiency with which the task is performed by the subject. The core task measure may or may not be directly comparable to these JPM measures of proficiency since proficiency is only captured by the performance of tasks beyond the core tasks and/or more difficult tasks than the core tasks.

Methods for Determining Core Tasks

One of the central issues which had to be addressed in the development of a productivity measure was how to determine which of the tasks listed in the OMS study are considered to be core tasks. The data element used in the selection of the core tasks was the participation rate for each of the tasks. For example, if one considers only first termers (48 months or less of active duty service and grade E4 or less) as in the JPM studies, then the core tasks can be identified as those tasks exhibiting a participation rate equal to or greater than a designated baseline participation rate. Two alternative methods were considered in establishing the baseline participation rate:

SME Method - This method involves the use of the distribution of participation rates by task (numerically ordered by size of the participation rate). The SME identifies a point along the distribution of the participation rates at which the slope last exhibits a significant increase in absolute value, such as the participation rate graphed in Figure 2 does at task 11. This point indicates that the participation rates for all tasks beyond this point are significantly lower than the participation rates of the tasks before this point. The core tasks are identified as those tasks with the higher participation rates prior to the change in the slope of the distribution.

E-value Method - This method is based on a mathematical algorithm which searches for the point along the distribution of participation rates (ordered higher to lower) where the

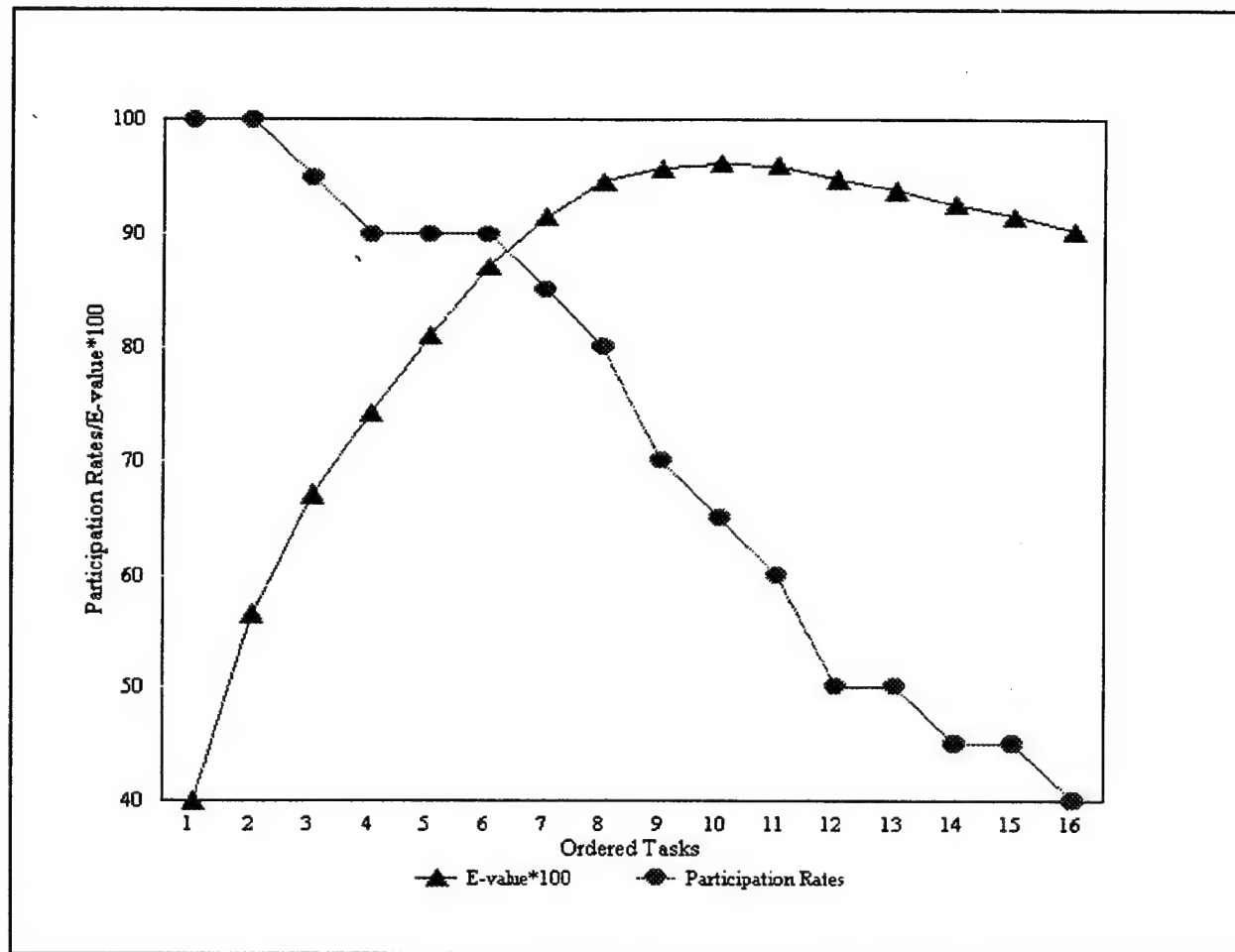


Figure 2. Comparison of Participation Rates and E-values

slope of the distribution begins decreasing (Stacey, Weissmuller, Barton, & Rogers, 1974). This is similar to a mathematical representation of the essence of the Subject Matter Expert Method. The calculation of an evaluation or E-value for a set of tasks is

$$EV_i = \frac{\sum_{i=1}^k x_i^2}{C\sqrt{i}} \quad (2)$$

where

- EV_i is the E-value rate for the i th iteration,
- i iteration number which represents the number of tasks in the cumulative sum of the denominator to that point,
- k is the total number of tasks,

- x is the participation rate for the task, where all tasks have been ordered high to low by participation rate, and
- C is a constant (25,000) for setting a standard against which to evaluate the level of participation rates used in computing the E-value function (Stacey et al., 1974).

Thus, the E-value method generates a vector of E-values, as specified in the above equation, for the tasks ordered highest to lowest and calculates a cumulative distribution of these E-values. The point along the cumulative distribution at which the E-values are less than or equal to the highest E-value for three consecutive E-values is determined. This point is the point at which the change in the participation rates has become insignificant. All tasks before and including the task with the highest obtained E-value are defined as the core set of tasks. The latest E-value (which includes the latest task) is selected if two or more of the E-values are tied for the highest E-value.

Table 2 provides an example of the E-value method of determining core tasks. The participation rates for the 16 tasks in Table 2 are ordered from highest to lowest participation rate (column 2); squared (column 3); a cumulative distribution is produced (column 4); E-value is calculated (column 5); and core tasks identified based on the values of column 5. In the example provided in Table 2, the first ten tasks would be identified as core tasks. Figure 2 demonstrates that task 10 represents the maximum point on the E-value cumulative distribution and occurs at a point in the distribution of participation rates where the level of participation in the tasks is falling significantly.

The SME method and the E-Value method were both used on OMS data for the eight JPM AFSs and compared for consistency. We present the results and compare both methods in the Estimation and Validation section. It should be noted at this point that the SME method was used to validate the results from the less-costly E-Value method. Now we turn our attention to the development of a productivity index to provide a measure to compare the two methods.

OMS Productivity Index

Although the Air Force occupational survey procedures do not directly measure productivity or proficiency, data elements are collected which lend themselves to the calculation of an index of productivity. The approach proposed in this paper uses the concept of core tasks to construct a measure or index reflecting whether an airman is less than, more than, or fully productive. This index will be greater than one if an individual is more than fully productive, less than one if an individual is less than fully productive, and equal to one if an individual is fully productive. The index uses two primary OMS data elements: (1) tasks performed by the respondent and (2) the difficulty index associated with the difficulty of learning how to perform the task.

Table 2. E-value Method Example

Task i	Rate	Rate²	Cumulative Rate²	E-value	Core Tasks
1	100	10000	10000	0.4000	*
2	100	10000	20000	0.5657	*
3	95	9025	29025	0.6703	*
4	90	8100	37125	0.7425	*
5	90	8100	45225	0.8090	*
6	90	8100	53325	0.8708	*
7	85	7225	60550	0.9154	*
8	80	6400	66950	0.9468	*
9	70	4900	71850	0.9580	*
10	65	4225	76075	0.9623	*
11	60	3600	79675	0.9609	
12	50	2500	82175	0.9489	
13	50	2500	84675	0.9394	
14	45	2025	86700	0.9269	
15	45	2025	88725	0.9163	
16	40	1600	90325	0.9033	

The first component in the OMS productivity index (OPI) is:

$$\frac{N_{c,j}}{N_c} \quad (3)$$

where

$N_{c,j}$ is the number of core tasks performed by the j th individual in the OMS sample, and

N_c is the number of core tasks.

The purpose of this component is to measure the proportion of the core tasks the individual is performing. If the individual is fully productive, that is performing all the core tasks, this component would be equal to one. If the individual is not performing all the core tasks, then he/she is less than fully proficient and this component will be less than one. The value of this component can vary between zero and one, but cannot exceed one. This component could be a productivity measure by itself, but this factor alone ignores the nature of tasks performed which could contribute to the productivity of an individual.

The second component of the OPI is defined as:

$$\frac{N_j}{N_c} \quad (4)$$

where

N_j is the total number of tasks performed by the j th individual in the OMS sample.

This component accounts for the total number of tasks performed by the individual compared to the number of core tasks. The component is greater than one if the individual is performing more tasks than the number of core tasks, and less than one if the individual is performing fewer tasks than the number of core tasks. This component does not account for the fact that the individual could be performing more tasks than the number of core tasks and still not be performing all the core tasks. Of course, if an individual is performing more tasks than the core tasks and not performing all the core tasks, this component will be greater than one but Equation 3 will be less than one. Only an individual performing more than the number of core tasks, as well as all the core tasks, will receive an average value of Equation 3 and Equation 4 of greater than one. An option to the calculation of this second component could be to use the ratio of the number of non-core tasks divided by the number of core tasks. This option would result in an individual receiving a value of zero for this component if he/she only performed the core tasks.

The third component of the OPI is calculated as:

$$\frac{\sum_{i=1}^t TD_{i,j}}{\sum_{i=1}^{t_c} TD_{i,c}} \quad (5)$$

where

$TD_{i,j}$ is the task difficulty associated with i th task performed by the j th individual and
 $TD_{i,c}$ is the task difficulty associated with i th task of the core tasks.

This component accounts for the fact that an individual could be assuming more difficult tasks at the same time that he/she is performing fewer core tasks. Thus, the ratio will be greater than or equal to one if an individual is performing more difficult tasks relative to the core tasks no longer being performed. The ratio would be greater than one if the individual is performing more tasks in addition to the core tasks.

Conversely, if an individual is not performing all the core tasks and is performing no additional non-core tasks (Table 3), this component will be less than one. If the individual is not performing all the core tasks, but is performing additional non-core tasks, the value of the component could be less than, equal to, or greater than one. If the additional non-core tasks being performed are less difficult relative to the core tasks not being performed (Table 3), the value for this component will be less than one. However, if the individual is performing enough non-core, less difficult tasks to offset the number and difficulty of core tasks not being performed (Table 3), the ratio will be equal to or greater than one. This allows the individual performing fewer than the number of core tasks to still potentially be fully productive, i.e. the value for this component would be equal to or greater than one.

Table 3. OPI Component Value Possibilities

Perform All Core Tasks	Perform Only Some of Core Tasks	Perform Tasks in Addition to Core Tasks	Possible Value of Third Component to OPI
X	--	No	= 1
X	--	X	> 1
--	X	No	< 1
--	X	X	< 1, = 1, or > 1*

* Value of third component of OPI is dependent upon the number and difficulty of core tasks not being performed, relative to the number and difficulty of additional tasks being performed.

The last component of the OPI is equal to the product of Equation 2 and Equation 5:

$$\left(\frac{\sum_{i=1}^t TD_{i,j}}{t_c} \times \frac{N_{c,j}}{N_c} \right) \frac{\sum_{i=1}^t TD_{i,c}}{\sum_{i=1}^t TD_{i,c}} \quad (6)$$

Equation 6 provides an additional penalty for not performing all of the core tasks. This component penalizes the individual for substituting non-core tasks for core tasks, as well as substituting less difficult non-core tasks for core tasks.

Calculation of the OMS Productivity Index (OPI)

Thus, the resulting equation for the calculation of OPI is

$$OPI_j = \frac{\frac{N_{c,j}}{N_c} + \frac{N_j}{N_c} + \frac{\sum_{i=1}^t TD_{i,j}}{t_c} + \left(\frac{\sum_{i=1}^t TD_{i,j}}{t_c} \times \frac{N_{c,j}}{N_c} \right) \frac{\sum_{i=1}^t TD_{i,c}}{\sum_{i=1}^t TD_{i,c}}}{4} \quad (7)$$

and each of the four components contributing to OPI are equally weighted in the calculation. Table 4 provides a summary definition of terms.

The OPI was developed for the purpose of providing a methodology for using OMS data to measure productivity. Simply using the number of core tasks performed did not account for all the possible combinations which could arise in the performance of tasks on the job. Equation 2 provides the simple productivity measure but does account for the performance of non-core tasks beyond or in place of core tasks. Thus, for those individuals who have mastered the core tasks and are being provided additional responsibilities, the OPI will reflect a more than fully productive individual. In addition, the OPI has also accounted for those individuals who trade non-core tasks for core tasks, as well as the difficulty of the non-core tasks compared to the difficulty of the core tasks.

Table 4. Summary of Definition of Terms

Term	Definition
OPI_j	Productivity index measure for j th individual based on OMS data.
$TD_{i,j}$	Task Difficulty associated with i th task performed by the j th individual.
$TD_{i,c}$	Task Difficulty associated with i th task of the core tasks.
N_j	Number of tasks performed by the j th individual
N_c	Number of core tasks.
$N_{c,j}$	Number of core tasks performed by the j th individual.

ESTIMATION AND VALIDATION OF PRODUCTIVITY INDEX USING EIGHT WTPT AFSs

Using the core task and OPI methodology described in the previous section, productivity functions were developed for the eight AFSs included in the Air Force JPM research program. The OPI functions were validated against productivity functions estimated using the JPM data. Additionally, the substitutability of the E-value method for determining core tasks for an AFS was tested against the SME method.

Variable and Sample Definitions

Productivity functions were estimated using OMS data collected for each of the eight AFSs from the Air Force JPM research program. The OMS data were selected to be chronologically closest to the administration of the JPM studies, but prior to the JPM study dates. In fact, these OMS studies should be the ones used by JPM researchers to select tasks for the WTPT. Additional data from the enlisted personnel master files were obtained for both OMS and JPM survey participants including: grade, TAFMS in months, AFQT scores, and ASVAB selector aptitude index (AI) scores.

Since the JPM samples were restricted to first-term airmen with fewer than 48 months of experience and a skill level of 3 or 5, the OMS samples were similarly restricted to facilitate comparison between the two data sources. In addition, the OMS samples were restricted to airmen whose number of tasks performed was within two standard deviations of the mean number of tasks performed by first-termers in the AFS, in order to remove airmen whose reported number of tasks performed was out of line with the rest of the airmen in the sample.

OMS sample sizes are provided in Table 5 for each AFS, along with the title for each of the AFSs. The restricted OMS samples were used to determine the core tasks for each AFS based upon the participation rates. Also provided in Table 5 for each of the AFSs are the appropriate selector AIs used for job classification. AFSs are classified using the four selector AIs: Mechanical (M), Administrative (A), General (G), or Electronic (E). The MAGE selector AIs are composites of selected subtest scores from the ASVAB.

Table 5. AFS Titles and OMS Sample Sizes

AFS - AI	Title	OMS Sample Size
122x0 - G	Aircrew Life Support	676
272x0 - G	Air Traffic Control	719
324x0 - E	Precision Measurement Equipment Laboratory	445
328x0 - E	Communication and Navigation Systems	406
423x5 - M	Aerospace Ground Equipment	1299
426x2 - M	Jet Engine Mechanic	1305
492x1 - A	Communication Systems Radio Operations	489
732x0 - A	Personnel	1730

Determination of Core Tasks

Two methods were used to determine the core tasks for each of the 8 AFSs, the SME method and the E-value method (discussed in the previous section). Table 6 shows the comparison of the number of tasks identified as core tasks using each method. Both methods produced similar results for all AFSs (Table 6) with the exception of the two M selector AI AFSs, 423x5 and 426x2. Both of these AFSs had very low participation rates compared to the other six AFSs, which may account for the differences in task selection between the two methods. OPI measures were computed for each AFS based on the identified core tasks from each method.

Estimation of OPI Equations

Using the methodology defined in the previous section, OPI was calculated for each airman in the OMS samples. One OPI was calculated using the E-value method for determining core tasks and a second OPI was calculated using the SME method. Table 7 contains descriptive

statistics for the OPI's calculated for each AFS using both methods. Mean values for the OPI vary by core task identification method for each AFS.

Table 6. Number of Core and Total Tasks

AFS - AI	Number of Core Tasks		Difference in Number of Core Tasks	Total Number of Core & Non-core Tasks
	SME Method	E-value Method		
122x0 - G	51	61	10	744
272x0 - G	54	65	11	485
324x0 - E	50	30	-20	1428
328x0 - E	58	58	0	1007
423x5 - M	34	80	46	615
426x2 - M	25	63	38	422
492x1 - A	33	31	-2	481
732x0 - A	20	15	-5	1541

For the two G selector AI AFSs (122x0 and 272x0), the mean value of the OPI calculated using the SME method for identifying core tasks was higher than the index using the E-value method. For both 122x0 and 272x0, the number of core tasks determined by the SME method was smaller than that of the E-value method, thus accounting for the higher mean values seen with the SME method. For the two A selector AI AFSs (492x1 and 732x0) the opposite occurred. Mean values of the OPI were higher using the E-value method, and the number of core tasks specified by the E-value method was also smaller than the number specified by the SME method. One of the E selector AI AFSs, 324x0, followed the same pattern as both of the A selector AI AFSs. For the other E selector AI AFS, 328x0, both methods of determining core tasks resulted in the same number and group of tasks, thus the value of the OPI was also the same under both methods. Large differences were seen in the means of the OPI for both of the M selector AI AFSs (423x5 and 426x2). Similar to the G AFSs (122x0 and 272x0), the mean value of the OPI using the SME method was higher than that using the E-value method. In the case of the M AFSs, however, the number of core tasks for each method varied greatly. For 423x5, 34 core tasks were identified using the SME method while 80 were identified as core tasks using the E-value method. For 426x2, 25 tasks were identified as core tasks using the SME method versus 63 core tasks using the E-value method. The large difference in the number of

core tasks by each method accounts for the large differences in the mean values of the OPI's for these two AFSs.

Table 7. Descriptive Statistics for the OPI by Core Task Identification Method

AFS - AI	Sample Size	SME Method Mean(St. Dev.)	E-value Method Mean(St. Dev.)
122x0 - G	676	1.55473 (0.964)	1.28435 (0.796)
272x0 - G	719	1.57132 (0.472)	1.29139 (0.388)
324x0 - E	445	1.63389 (1.059)	2.92018 (1.903)
328x0 - E	406	1.44469 (0.762)	1.44469 (0.762)
423x5 - M	1299	2.20874 (1.685)	0.91249 (0.693)
426x2 - M	1305	1.68019 (1.110)	0.67676 (0.436)
492x1 - A	489	2.23423 (1.091)	1.57095 (0.760)
732x0 - A	1730	1.70655 (1.222)	2.37593 (1.711)

For both core task methods the productivity/proficiency (OPI) relationship with aptitude and experience was estimated using the following ordinary least squares equation:

$$OPI = f(\text{experience, aptitude, bases})$$

or

$$OPI = \beta_0 + \beta_1 E + \beta_2 E^2 + \beta_3 A + \sum \beta_j Base_j + \epsilon \quad (8)$$

where

- E* is experience defined using TAFMS in months,
A is aptitude defined using the appropriate selector AI (M, A, G or E) for each AFS, and
Base is the binary variable for the base at which the airman is located.

Previous research using JPM data for the Air Force has shown substantial evidence for a non-linear relationship between proficiency and experience (Stone et al., 1991). Therefore, the non-linear experience relationship was used in the OPI estimations. The appropriate selector AI for each of the eight AFSs is presented in Table 8. Table 8 also contains the mean and standard deviation for the values of TAFMS and aptitude for each AFSs, as well as the number of airman in each AFS. Binary variables were included to account for any variation in productivity/proficiency due to the base to which the airman was assigned while being tested.

Table 8. Descriptive Statistics for TAFMS and Aptitude from OMS Sample

AFS - AI	Sample Size	TAFMS Mean(St. Dev.)	Aptitude Mean(St. Dev.)
122x0 - G	676	28.99 (11.4)	54.15 (19.2)
272x0 - G	719	28.06 (11.2)	73.08 (16.0)
324x0 - E	445	26.02 (8.8)	79.76 (10.7)
328x0 - E	406	30.19 (11.0)	87.09 (5.3)
423x5 - M	1299	27.11 (10.8)	54.48 (20.5)
426x2 - M	1305	26.46 (11.9)	69.78 (18.1)
492x1 - A	489	24.29 (11.4)	71.49 (14.2)
732x0 - A	1730	27.56 (11.4)	73.52 (15.8)

Table 9 presents the estimation results using the E-value method and Table 10 presents the estimation results using the SME method. Coefficient values and t-statistics are provided in the tables for the experience and aptitude variables in the equations. Coefficients for the binary variables accounting for differences between bases are included in the constant term. As Table 9 shows, most variables were significant at the 90% level of significance or greater. For both 122x0 and 272x0 (G selector AI AFSs), both aptitude and experience were significant at least the 95% level of significance using either method. The R^2 s and t-statistics using either method for determining core tasks were also similar. Figure 3 presents the productivity function derived from the coefficient estimates in Table 9 using the E-value method for determining core tasks for AFS 122x0.

Table 9. OPI Estimation Results -- E-value Method

AFS - AI	TAFMS	TAFMS ²	Aptitude	Constant	R-square/obs
122x0 - G	0.06014 (4.947)*	-0.00090 (-4.254)*	0.00382 (2.762)*	0.20542 (0.099)	0.4346 676
272x0 - G	0.05189 (6.741)*	-0.00068 (-5.228)*	0.00175 (2.018)**	0.32425 (2.185)**	0.3320 719
324x0 - E	0.31014 (4.578)*	-0.00435 (-3.834)*	0.01544 (1.910)***	-3.10295 (-4.053)*	0.3877 445
328x0 - E	0.00682 (0.353)	-0.00002 (-0.072)	0.00210 (0.311)	1.08117 (1.826)***	0.4323 406
423x5 - M	0.22193 (2.309)**	-0.00031 (-1.817)***	0.00526 (5.662)*	0.28416 (1.332)	0.1993 1299
426x2 - M	0.02494 (4.451)*	-0.00034 (-3.389)*	0.00285 (4.360)*	0.10338 (1.339)	0.1820 1305
492x1 - A	0.08570 (5.188)*	-0.00118 (-4.056)*	-0.00031 (-0.137)	0.36330 (0.475)	0.3680 489
732x0 - A	0.07000 (3.830)*	-0.00072 (-2.192)**	0.00883 (3.481)*	0.43517 (-0.160)	0.2011 1730

* - $p < .01$
 ** - $p < .05$
 *** - $p < .10$

Table 10. OPI Estimation Results -- SME Method

AFS - AI	TAFMS	TAFMS ²	Aptitude	Constant	R-square/obs
122x0 - G	0.07293 (4.950)*	-0.00109 (-4.262)*	0.00467 (2.782)*	0.24582 (0.064)	0.4345 676
272x0 - G	0.63974 (6.835)*	-0.00084 (-5.329)*	0.00220 (2.086)**	0.37988 (0.034)	0.3354 719
324x0 - E	0.16769 (4.466)*	-0.00237 (-3.771)*	0.00810 (1.808)***	-1.58862 (-3.894)*	0.3924 445
328x0 - E	0.00682 (0.353)	-0.00002 (-0.072)	0.00210 (0.311)	1.08117 (1.826)***	0.4323 406
423x5 - M	0.05316 (2.274)**	-0.00073 (-1.785)***	0.01247 (5.520)*	0.71116 (1.362)	0.1996 1299
426x2 - M	0.06357 (4.460)*	-0.00084 (-3.299)*	0.00747 (4.493)*	0.18383 (1.118)	0.1826 1305
492x1 - A	0.07993 (5.160)*	-0.00110 (-4.030)*	-0.00030 (-0.141)	0.34893 (0.495)	0.3690 489
732x0 - A	0.04865 (3.730)*	-0.00049 (-2.100)**	0.00623 (3.440)*	0.34382 (-0.032)	0.2016 1730

* - $p < .01$

** - $p < .05$

*** - $p < .10$

For AFS 324x0, experience was significant at the 99% level of confidence, while aptitude was only marginally significant using either method for determining core tasks. As was seen in Table 8, the mean for aptitude for AFS 324x0 was 79.76 with a standard deviation of 10.7, suggesting minimal variation in the aptitude variable, since the aptitude distribution has been severely truncated at the higher aptitude levels. Results using both core tasks methods were the same, as expected, for AFS 328x0. The OPI equation results were weak for AFS 328x0; neither aptitude nor experience were found to have a statistically significant effect on OPI. Once again, the aptitude distribution was severely truncated at the higher aptitudes levels for an E selector AI AFS, 328x0 with a mean value of 87.09 and a standard deviation of 5.3.

For both AFSs 423x5 and 426x2, all variables were statistically significant at the 95 percent level of confidence, with the exception of the squared experience term (TAFMS²) for AFS 423x5 using either method to determine the number of core tasks. R²s for these two AFSs were relatively low.

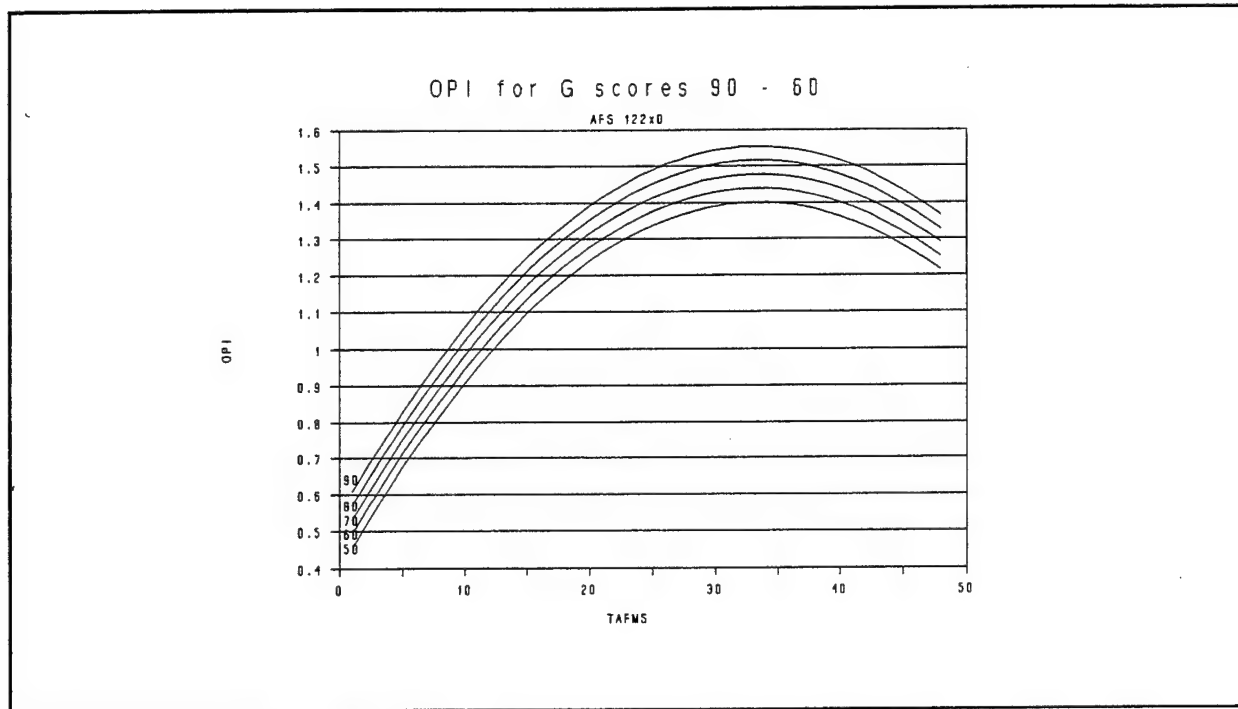


Figure 3. Example Productivity Function Using OPI

Coefficient Comparison Between E-Value and SME Method

In order to compare coefficients between the two core task identification methods, the equations for each method were re-estimated using OPI values which had been standardized (mean = 0, standard deviation = 1) by AFS. The re-estimations yielded standardized coefficients which were comparable between the two core task identification methods.

The standardized coefficients are presented in Table 11. Differences between the coefficients of each core task method for the experience and aptitude variables in the equations are also shown in the table. A simple means test (Mendenhall & Scheaffer, 1973) was performed to determine if the coefficients between the E-value method and the SME method were statistically different from one another. The t-values for this test are also provided in Table 11.

As can be seen in Table 11, only a few variables proved to have coefficients that were statistically different between the two core task identification methods at the 99% level of confidence. The coefficients for aptitude for AFS 423x5 proved to be statistically different between the two core task identification methods. The coefficients for TAFMS² and aptitude for AFS 426x2 also proved to be statistically different between the two methods. The coefficients for TAFMS and TAFMS² for AFS 732x0 were statistically different between the two core task identification methods. The three AFSs which exhibited coefficients that were

Table 11. Comparison of Standardized Coefficients for SME and E-value Methods

AFS - AI	Variable	Coefficient SME method	Coefficient E-value method	% Difference	Means Test t-statistic
122x0 - G	TAFMS	0.0756	0.0756	0.08	0.076
	TAFMS ²	-0.0011	-0.0011	0.18	-0.138
	Aptitude	0.0048	0.0048	0.73	0.375
272x0 - G	TAFMS	0.1354	0.1339	1.13	1.459
	TAFMS ²	-0.0018	-0.0017	1.65	-1.667
	Aptitude	0.0047	0.0045	3.02	1.194
324x0 - E	TAFMS	0.1584	0.1630	-2.90	-1.925
	TAFMS ²	-0.0022	-0.0023	-2.06	1.159
	Aptitude	0.0077	0.0081	-6.03	-1.623
328x0 - E	TAFMS	0.0089	0.0089	-----	-----
	TAFMS ²	-0.0000	-0.0000	-----	-----
	Aptitude	0.0028	0.0028	-----	-----
423x5 - M	TAFMS	0.0315	0.0320	-1.53	-0.884
	TAFMS ²	-0.0004	-0.0004	-1.83	0.831
	Aptitude	0.0074	0.0076	-2.60	-3.654*
426x2 - M	TAFMS	0.0573	0.0572	0.16	0.178
	TAFMS ²	-0.0008	-0.0008	-2.77	2.330*
	Aptitude	0.0067	0.0065	2.92	3.348*
492x1 - A	TAFMS	0.1121	0.1128	-0.62	-0.498
	TAFMS ²	-0.0015	-0.0016	-0.72	0.456
	Aptitude	-0.0004	-0.0004	2.86	0.063
732x0 - A	TAFMS	0.0398	0.0409	-2.70	-2.965*
	TAFMS ²	-0.0004	-0.0004	-4.41	2.722*
	Aptitude	0.0051	0.0052	-1.22	-1.234

* - $p < .01$

statistically different between core tasks identification methods also exhibited relatively low R-squares (Tables 9 and 10).

Comparison of the coefficients between the two core tasks identification methods suggests that in the majority of AFSs, the two methods are interchangeable. Both tend to produce coefficients which cannot be statistically differentiated from one another. The AFSs which produced coefficients that were statistically different had a great variety in the number and type of tasks performed by airmen. For example, the number and type of tasks performed by airmen in AFS 732x0 varies widely compared to other AFSs. The same is true of AFSs 423x5 and 426x2 with tasks which vary by base and type of aircraft. All three of these AFSs (423x5, 426x2, and 732x0) also displayed relatively low R-squares when compared with the other AFSs. Therefore, the E-value method of identifying core tasks can be used as a surrogate for the more costly SME method without appreciably affecting the quality of the estimated OPI relationship with aptitude and experience.

Estimation of WTPT Proficiency Equation

In order to validate the OPI as a measure of productivity/proficiency, an alternate measure of proficiency from the JPM WTPT data was used to estimate an aptitude/experience equation. For each of the eight AFSs, an average proficiency (APF) measure was calculated for each airman using the information gathered in the WTPT on the number of steps performed correctly across tasks. The APF was calculated as the grand mean of steps performed correctly by task, averaged across all tasks performed by each airman in the WTPT. Table 12 shows the distribution of the APF measure for the eight AFSs, as well as the number of airmen in each AFS for the JPM sample.

Table 12. Descriptive Statistics for APF in the JPM Sample

AFS - AI	Sample Size	APF Mean	APF St. Dev.
122x0 - G	171	0.66523	0.133
272x0 - G	172	0.66659	0.100
324x0 - E	124	0.77764	0.080
328x0 - E	67	0.73698	0.121
423x5 - M	218	0.52640	0.096
426x2 - M	197	0.71372	0.113
492x1 - A	123	0.77691	0.140
732x0 - A	175	0.79885	0.111

Table 13 contains the means and standard deviations for TAFMS and aptitude for each AFS, as well as the number of airmen in each JPM AFS. Of importance to note is the aptitude distribution of AFSs 324x0 and 328x0. The mean aptitude score for these two AFSs is approximately 85, considerably higher than any of the other AFSs, and the standard deviation is considerably smaller than any of the other AFSs. This lack of variation in aptitude could reduce the quality of the statistical results in estimating the relationship between proficiency and aptitude. These two AFSs also had relatively small sample sizes of airmen tested.

Table 13. Descriptive Statistics for TAFMS and Aptitude in the JPM Sample

AFS - AI	Sample Size	TAFMS Mean(St. Dev.)	Aptitude Mean(St.Dev.)
122x0 - G	171	28.52 (10.9)	59.71 (17.5)
272x0 - G	172	26.89 (8.8)	74.74 (14.7)
324x0 - E	124	26.40 (10.3)	84.25 (9.6)
328x0 - E	67	30.14 (11.4)	85.18 (9.3)
423x5 - M	218	27.99 (10.3)	76.01 (13.2)
426x2 - M	197	28.97 (10.7)	76.65 (16.1)
492x1 - A	123	23.49 (12.8)	70.78 (14.3)
732x0 - A	175	26.74 (11.0)	73.14 (14.7)

An APF aptitude/experience relationship was estimated using the following ordinary least squares specification:

$$\text{APF} = f(\text{experience, aptitude, bases, raters})$$

or

$$APF = \alpha_0 + \alpha_1 E + \alpha_2 E^2 + \alpha_3 A + \sum \alpha_j Base_j + \sum \alpha_k Rater_k + \epsilon \quad (9)$$

where

- E** is experience defined using TAFMS in months,
- A** is aptitude defined using the appropriate selector AI (M, A, G or E) for each AFS,
- Base** is the binary variable for the base at which the airman is located, and
- Rater** is the binary variable for the rater evaluating the airman.

Previous research using JPM data for the Air Force has shown substantial evidence for a non-linear relationship between proficiency and experience (Stone et al, 1992). Therefore, the non-linear experience relationship was used in the APF estimations. Binary variables were included to account for any variation in proficiency due to the base the airman was assigned or the rater evaluating the airman.

Table 14 presents the estimation results using the APF relationship specified in Equation 9. Coefficient values and t-statistics are shown in the tables for experience and aptitude variables in the equations. Coefficients for the binary variables accounting for differences between bases and raters have been included in the constant. As Table 14 shows, the experience and aptitude coefficients did not demonstrate any consistency in statistical significance across AFSs as was seen in Table 9 for the OPI relationships. For 272x0, both aptitude and experience were statistically significant at the 90 percent level of significance or greater. No other APF equation in Table 14 had statistically significant coefficients for aptitude, experience, and experience squared. Two other APF equations had statistically significant coefficients for aptitude and experience, 324x0 and 423x5, though the TAFMS² in both equations was statistically insignificant.

For several of the AFSs in Table 14, the results suggest that bases and raters explain more of the variation in the APF dependent variable than do aptitude and experience. Aptitude was statistically significant in five of the eight AFSs of Table 14 (four at the 99 percent level of confidence and one at the 90 percent level of confidence). TAFMS was statistically significant in four of the eight AFSs in Table 14 (one at the 99 percent level of confidence, two at the 95 percent level of confidence, and one at the 90 percent level of confidence). TAFMS in Table 9 was statistically significant in seven of the eight AFSs (six at the 99 percent level of confidence and one at the 90 percent level of confidence). The patterns of statistical significance by AFS between Tables 9 and 14 are also inconsistent. APF equations which exhibited statistically significant coefficients in Table 14 were quite similar to the coefficients of Table 9 (such as AFSs 272x0, 324x0, 423x5, and 492x1). The APF equations which exhibited poor coefficients (no statistical significance) were totally different (such as AFSs 122x0, 426x2, and

732x0). In general, OPI equations presented in Table 9 provided statistically superior results to the APF equations in Table 14.

Table 14. APF Estimation Results

AFS - AI	TAFMS	TAFMS ²	Aptitude	Constant	R-square/obs
122x0 - G	0.00317 (0.689)	0.00000 (0.043)	0.00031 (0.572)	0.55321 (5.877)*	0.4593 171
272x0 - G	0.01007 (2.131)**	-0.00015 (-1.876)***	0.00228 (5.356)*	0.34791 (4.857)*	0.5168 172
324x0 - E	0.00897 (1.955)***	-0.00011 (-1.437)	0.00254 (3.441)*	0.41738 (4.356)*	0.4791 124
328x0 - E	0.01088 (0.910)	-0.00013 (-0.672)	0.00270 (1.879)***	0.31932 (1.110)	0.6825 67
423x5 - M	0.00650 (1.946)**	-0.00008 (-1.347)	0.00270 (6.454)*	0.20710 (3.868)*	0.4501 218
426x2 - M	0.00689 (1.601)	-0.00009 (-1.233)	0.00156 (3.856)*	0.47821 (4.109)*	0.5272 197
492x1 - A	0.01479 (3.867)*	-0.00021 (-3.001)*	0.00073 (1.215)	0.52574 (7.046)*	0.6737 123
732x0 - A	0.00498 (1.411)	-0.00002 (-0.392)	0.00065 (1.332)	0.63916 (5.022)*	0.4784 175

* - $p < .01$

** - $p < .05$

*** - $p < .10$

Comparative Statistics (Correlations, Spearman Rank Order, Kendall's-tau)

Since APF is considered to be a measure of proficiency with respect to the performance of tasks within AFSs, comparisons were made between APF and OPI using the JPM and OMS data for each AFS. These comparisons focused on the predictive similarities of the APF and OPI equations based on aptitude and experience since these equations control for variations in these dependent variables which are beyond the effects of aptitude and experience. Several tests were used to determine the level of comparability between the OPI equations and the APF equations with respect to the prediction of the productivity/proficiency of airmen. The test statistics used were simple correlations, Spearman rank order correlations, and Kendall's-tau statistic.

In order to test the comparability of the predictive strengths of the OPI and APF equations, both the APF equation and the OPI equation were used to predict the proficiency of airmen in the JPM sample for each AFS. First, the OPI coefficient estimates presented in Tables 8 and 9 for each of the core task methods (E-value and SME), were applied to airmen in the JPM sample. The OPI equation coefficients were used to obtain the predicted OPI productivity/proficiency measure for each airman for each of the core task identification methods. The predicted OPI for the JPM sample was obtained using the average probability (i.e., sample mean of the binary base variable from the OMS sample) of being assigned to a base for each of the bases in the equations.

Next, the APF coefficient estimates presented in Table 14 were applied to the JPM sample. To obtain the predicted APF value, aptitude and experience (TAFMS and TAFMS²) were set to zero for each airman in the JPM sample and the resulting predicted value was obtained. This predicted value represented the systematic variation in APF caused by base and raters. The residual predicted APF value was then calculated by subtracting the predicted APF value from the actual APF value for each airman. This residual represented the variation in APF caused by aptitude, experience, and error. This residual APF value was then compared against the predicted OPI value for each airman.

The comparisons of the predicted values for OPI and the residual APF are presented in Table 15 and Table 16. Table 15 presents the results using the E-value method, and Table 16 presents the results using the SME method. The Kendall's-tau, Spearman Rank Order Correlation, and simple correlation statistics are shown for each of the eight AFSs. The last column in each of the tables presents the correlation between the predicted OPI values for each core task identification method and the actual APF scores. Values for the Kendall's-tau statistic were consistently lower than for the Spearman rank order correlations, but all were statistically significant at the 99% level of confidence. Values for the Spearman rank order correlation ranged from 0.2594 to 0.5666, suggesting that the OPI equation using the E-value method rank-ordered airmen similarly to the APF equation. In each case, the correlation between the residual APF values and the predicted OPI values was higher than the actual APF values and the predicted OPI values. Values for the simple correlation between the residual APF and the predicted OPI values using the E-value method ranged from a low of 0.2641 to a high of 0.5524. Similar results were seen for the predicted OPI values using the SME method.

Magnitudinal Effect of Experience and Aptitude

Table 17 presents the standardized coefficients for both the JPM and the OPI/E-Value Method productivity equations. These coefficients allow direct comparison of the magnitudinal effect of experience and aptitude as estimated from JPM versus OMS based productivity equations. Comparisons can only be made between coefficients which were statistically significant in both equations.

Table 15. Comparison Statistics -- E-value Method

AFS - AI	Kendall's - tau Statistic	Spearman Rank Order Correlation	Simple Correlation	Correlation with Actual APF
122x0 - G	0.1760*	0.2594	0.3083	0.2218
272x0 - G	0.1843*	0.2709	0.2641	0.2147
324x0 - E	0.3469*	0.5076	0.5083	0.4481
328x0 - E	0.3965*	0.5666	0.5098	0.2007
423x5 - M	0.3547*	0.5005	0.5080	0.4064
426x2 - M	0.2330*	0.3388	0.3447	0.1432
492x1 - A	0.3507*	0.5009	0.5524	0.5420
732x0 - A	0.3104*	0.4561	0.4432	0.4301

* - $p < .01$

Table 16. Comparison Statistics -- SME Method

AFS - AI	Kendall's - tau Statistic	Spearman Rank Order Correlation	Simple Correlation	Correlation with Actual APF
122x0 - G	0.1761*	0.2595	0.3079	0.2214
272x0 - G	0.1880*	0.2757	0.2673	0.2169
324x0 - E	0.3425*	0.5011	0.5062	0.4449
328x0 - E	0.3965*	0.5666	0.5098	0.2007
423x5 - M	0.3548*	0.5006	0.5077	0.4063
426x2 - M	0.2330*	0.3370	0.3429	0.1402
492x1 - A	0.3515*	0.5023	0.5524	0.5419
732x0 - A	0.3108*	0.4565	0.4439	0.4296

* - $p < .01$

Table 17. Comparison of Standardized Coefficients for JPM and E-value Method

AFS - AI	Variable	Coefficient JPM	Coefficient E-value method	% Difference	Means Test T-statistic
122x0 - G	Experience	-----	0.0118	----	-----
	Aptitude	-----	0.0048	----	-----
272x0 - G	Experience	0.0203	0.0385	-89.66	-----
	Aptitude	0.0228	0.0045	80.16	78.377*
324x0 - E	Experience	0.1125 ⁺	0.0433	----	-----
	Aptitude	0.0318	0.0081	74.51	40.846*
328x0 - E	Experience	-----	-----	----	-----
	Aptitude	0.0222	-----	----	-----
423x5 - M	Experience	0.0679 ⁺	0.0103	----	-----
	Aptitude	0.0282	0.0076	73.06	136.104*
426x2 - M	Experience	-----	0.0149	----	-----
	Aptitude	0.0138	0.0065	52.81	50.155*
492x1 - A	Experience	0.0355	0.0351	1.21	-----
	Aptitude	-----	-----	----	-----
732x0 - A	Experience	-----	0.0189	----	-----
	Aptitude	-----	0.0052	----	-----

^{*} - $p < .01$

⁺ - Includes coefficient for TAFMS only.

Aptitude was statistically significant in both equations for three of the eight JPM AFSs, 272x0, 324x0, and 423x5. In each of these three cases, the coefficients for aptitude were statistically different (statistically significant means test) from each other at the 99 percent level of confidence. In addition, the estimated effect for aptitude was larger for the JPM equations versus the OMS counterparts in all three cases. In fact, the mean value for the aptitude coefficient from the five JPM equations in which aptitude was statistically significant was 0.0256 with a standard deviation of 0.0068. Thus, the statistically significant aptitude coefficients for the five JPM equations were all of similar magnitude. The same can be inferred from the six OMS equations in which the aptitude coefficient was statistically significant, with a mean of

0.0061 with a standard deviation of 0.0014. Thus, the statistically significant coefficients for aptitude for the JPM and OMS productivity equations were within similar magnitudinal ranges within the JPM or OMS equations, but ranges between the two equations were magnitudinally different.

Experience provides different results compared to aptitude. The calculation of the change in productivity for a unit change in experience is slightly complicated since experience is modeled as a nonlinear effect. The change in productivity for a unit change in experience, or the first derivative of productivity with respect to experience, can be expressed as

$$\frac{\partial OPI}{\partial E} = \beta_1 + 2\beta_2 E \quad (10)$$

for OPI and

$$\frac{\partial APF}{\partial E} = \alpha_1 + 2\alpha_2 E \quad (11)$$

for APF. The values presented in Table 17 for experience were evaluated at the mean value for experience using equations 10 and 11. Experience was statistically significant in only two of the eight JPM AFSs, 272x0 and 492x1 (both TAFMS and TAFMS² statistically significant). The statistically significant coefficients for JPM and OMS productivity equations are not magnitudinally different from each other for AFS 492x1 (1.2 percent) while significantly different for AFS 272x0 (89.7 percent). The mean value for the statistically significant experience coefficients (both TAFMS and TAFMS² are statistically significant) for the seven OMS equations was 0.0247 with a standard deviation of 0.0128. The mean value for the statistically significant experience coefficients for the two JPM equations was 0.0279 with a standard deviation of 0.0077. Thus, the statistically significant coefficients for experience for the JPM and OMS productivity equations were within similar magnitudinal ranges both within and across the JPM and OMS equations.

EXTENSION TO OTHER AFSs

Having established the ability of the OPI equations to measure and predict the productivity of individual airmen, the OPI methodology was extended to a wider set of specialties or occupational clusters. In extending the methodology, the enlisted career field structure was divided into 17 clusters of AFSs. These clusters were created based on the authors' judgement of similarities among specialties at the 2-digit classification level. One AFS was chosen from each of the 17 clusters to be a representative AFS for each cluster. Table 18 presents the 2-digit job classification clusters, as well as the representative AFS selected for each cluster along with the selector AI for that AFS.

Table 18. Definition of Clusters

Cluster	2-digit Cluster grouping	AFS	Selector AI
1	41xx Systems Maint 46xx Munitions & Weapons	461x0	M
2	49xx Commun Computer Systems	492x1	A
3	45xx Manned Aero Maint	426x2	M
4	32xx Precision Measurement 40xx Intricate Equip Maint	324x0	E
5	24xx Systems Maint 57xx Fire Protection 81xx SP	811x0	G
6	90xx Medical 91xx Medical 92xx Medical 98xx Dental	902x0	G
7	39xx Main Mgt Systems 60xx Transportation 61xx Commissary 62xx Services 63xx Fuels 64xx Supply	603x0	M
8	22xx Geodetic 23xx Visual Info	231x0	G
9	27xx Command Ctrl Sys Opr 30xx Comm-Elec Systems	272x0	G
10	36xx Wire Comm Systems Maint 54xx Mech/Electrical 55xx Structural/Pavements 56xx Sanitation	542x0	E
11	47xx Vehicle Maint 59xx Marine	472x2	M
12	31xx Instrumentation 34xx Training Devices	316x3	E

Table 18. Continued

Cluster	2-digit Cluster grouping	AFS	Selector AI
13	11xx Aircrew Ops 12xx Aircraft Protection	122x0	G
14	73xx Personnel 76xx Public Affairs	732x0	A
15	20xx Intelligence 82xx Special Investigation 88xx Paralegal	201x0	G
16	10xx First Sgt 25xx Weather 66xx Logistics Plans 70xx Information Management 74xx MWR 75xx Education and Training 87xx Band 89xx Chapel Management	702x0	A
17	65xx Contracting 67xx Financial	651x0	G

For the 17 representative AFSs selected, OPI productivity functions were estimated using the methodology defined in the previous sections. Productivity functions were estimated using OMS data for the AFSs. The latest OMS study available for each AFS was used in the estimation. Data from the enlisted personnel master files were again obtained for each OMS survey participant. These data again included: grade, TAFMS date, AFQT scores, and ASVAB selector AI scores. For all tasks included in the OMS surveys, task difficulty indexes were also obtained.

The OMS samples were again restricted to first-term airmen with less than 48 months of experience and skill level 3 or 5. In addition, the samples were restricted to airmen performing a number of tasks within two standard deviations of the mean number of tasks performed for that AFS. The restricted OMS samples were used to determine the core tasks for each AFS based upon participation rates. The E-value method was used to determine the core tasks for each of the AFSs (Table 19).

Based upon the core tasks defined using the E-value method, OPI was calculated for each airman in each of the AFSs. Mean values and standard deviations for OPI are presented in Table 20. Mean values ranged from a low of 0.677 for AFS 426x2 to a high of 2.920 for AFS

Table 19. Core Tasks

AFS - AI	Number of Core Tasks	Total Number of Tasks
461x0 - M	34	623
492x1 - A	31	481
426x2 - M	63	422
324x0 - E	30	1428
811x0 - G	24	886
902x0 - G	76	916
603x0 - M	28	367
231x0 - G	70	200
272x0 - G	65	485
542x0 - E	45	550
472x2 - M	132	1633
316x3 - E	26	877
122x0 - G	61	744
732x0 - A	15	1541
201x0 - G	33	848
702x0 - A	33	969
651x0 - G	14	1363

324x0. The productivity/proficiency (OPI) relationship with aptitude and experience was estimated using ordinary least squares with the same specification as Equation (8). The non-linear experience relationship was used again in the OPI estimations. TAFMS in months was again used for experience, and the appropriate selector AI (M, A, G, or E) for each AFS was used for aptitude. Table 21 presents the means and standard deviations for the values of TAFMS and aptitude for each AFS, as well as the number of airmen in each OMS sample. Binary variables were included to account for any variation in proficiency due to the base to which the airman was assigned.

Table 20. Descriptive Statistics for OPI in the Cluster AFSs

AFS - AI	Sample Size	OPI Mean	OPI St. Dev.
461x0 - M	1757	1.19442	0.741
492x1 - A	489	1.57095	0.760
426x2 - M	1305	0.67676	0.436
324x0 - E	445	2.92018	1.903
811x0 - G	1340	1.47382	0.971
902x0 - G	1382	1.08035	0.574
603x0 - M	1265	1.53139	0.909
231x0 - G	70	0.70628	0.472
272x0 - G	719	1.29239	0.388
542x0 - E	428	1.28538	0.802
472x2 - M	309	1.09488	0.697
316x3 - E	49	1.57213	1.01
122x0 - G	676	1.28435	0.796
732x0 - A	1730	2.37593	1.711
201x0 - G	200	1.17246	0.880
702x0 - A	894	1.45536	1.014
651x0 - G	145	1.89713	1.489

Table 22 presents the estimation results. Coefficient values and t-statistics are shown for the experience and aptitude variables in the equations. Coefficients for the binary variables accounting for differences between bases have been included in the constant term. For most of the AFSs, experience and aptitude were significant at the 90 percent level of confidence or greater. R-squares varied from a low of 0.0253 for AFS 461x0 to a high of 0.8761 for AFS 231x0. No variables (other than bases) for AFS 316x3 were found to be significant at greater than the 90 percent level of confidence, but this AFS was hampered by its small sample size (only 49 first-term survey participants). Limited success was seen with AFS 702x0; only TAFMS was significant at the 95 percent level of confidence. The limited success of AFS 702x0 can also be partially attributed to the wide variety of tasks performed across airman in

Table 21. Descriptive Statistics for Experience and Aptitude in the Cluster AFSs

AFS - AI	Sample Size	Experience Mean(St. Dev.)	Aptitude Mean(St. Dev.)
461x0 - M	1757	28.09 (11.9)	71.18 (17.5)
492x1 - A	489	23.49 (12.8)	70.78 (14.3)
426x2 - M	1305	28.97 (10.7)	74.74 (14.7)
324x0 - E	445	26.40 (10.3)	84.25 (9.6)
811x0 - G	1340	29.16 (12.1)	62.06 (15.3)
902x0 - G	1382	24.51 (11.0)	66.95 (15.4)
603x0 - M	1265	27.53 (11.4)	69.74 (15.2)
231x0 - G	70	33.09 (11.5)	74.86 (12.6)
272x0 - G	719	26.89 (8.8)	74.74 (14.7)
542x0 - E	428	22.59 (11.5)	64.89 (17.3)
472x2 - M	309	25.72 (14.7)	77.35 (12.8)
316x3 - E	49	28.90 (8.0)	80.49 (9.2)
122x0 - G	676	28.52 (10.9)	59.71 (17.5)
732x0 - A	1730	26.74 (11.0)	73.14 (14.7)

Table 21. Continued

AFS - AI	Sample Size	Experience Mean(St. Dev.)	Aptitude Mean(St. Dev.)
201x0 - G	200	19.85 (10.7)	72.88 (15.4)
702x0 - A	894	27.87 (12.2)	68.80 (16.6)
651x0 - G	145	27.57 (10.9)	69.12 (18.7)

the career field which makes determining core tasks for this AFS very difficult (similar to AFS 732x0 in the previous section).

For most AFSs, however, experience and aptitude were important factors in explaining the variation in the OPI index for airman in the career field. TAFMS was statistically significant in 14 AFSs; 12 at the 99 percent level of confidence and two at the 95 percent level of confidence. TAFMS² was statistically significant in 13 of the 17 AFSs; 10 at the 99 percent level of confidence, two at the 95 percent level of confidence, and one at the 90 percent level of confidence. Aptitude was statistically significant in 10 AFSs; six at the 99 percent level of confidence, two at the 95 percent level of confidence, and two at the 90 percent level of confidence. In eight AFSs, aptitude, experience, and experience squared were statistically significant at the 90 percent level of confidence or better. In all the other AFSs, with the exception of 316x3, either aptitude or experience (experience squared) was statistically significant at the 90 percent level of confidence or better.

Table 23 presents the standardized coefficients for the OMS AFSs. The statistically significant coefficients for aptitude display a large magnitudinal range, from a low of 0.0035 for AFS 902x0 to a high of 0.0208 for AFS 472x0. AFSs 472x2 and 651x0 exhibit magnitudinally high aptitude coefficients compared to the other eight statistically significant coefficients which exhibit a mean of 0.0054 with AFSs 472x2 and 651x0 excluded from the calculation versus a mean of 0.0077 (42.6 percent higher) with AFSs 472x2 and 651x0 included in the calculation.

The magnitudinal effect for experience also displays a wide range of variation, from a high of 0.2709 for AFS 231x0 to a low of 0.0076 for AFS 811x0. The high value for AFS 231x0 is significantly higher than any of the other 13 statistically significant coefficients which exhibit a mean of 0.0230 with AFS 231x0 excluded from the calculation versus a mean of 0.0407 (76.96 percent higher) with AFS 231x0 included in the calculation.

Table 22. OPI Estimation Results for the Cluster AFSs

AFS - AI	TAFMS	TAFMS ²	Aptitude	Constant	R-square/obs
461x0 - M	0.02692 (3.946)*	-0.00035 (-2.822)*	0.00341 (3.389)*	0.52072 (4.493)*	0.0253 1757
492x1 - A	0.08570 (5.188)*	-0.00118 (-4.056)*	-0.00031 (-0.137)	0.36330 (0.475)	0.3680 489
426x2 - M	0.02494 (4.451)*	-0.00034 (-3.389)*	0.00285 (4.360)*	0.10338 (1.339)	0.1820 1305
324x0 - E	0.31014 (4.578)*	-0.00435 (-3.834)*	0.01544 (1.910)***	-3.10295 (-4.053)*	0.3877 445
811x0 - G	0.04700 (3.675)*	-0.00070 (-3.175)*	0.00559 (3.227)*	0.45547 (1.879)***	0.1867 1340
902x0 - G	0.01764 (2.406)**	-0.00022 (-1.659)***	0.00201 (1.936)***	0.67465 (4.010)*	0.1194 1382
603x0 - M	0.04479 (4.322)*	-0.00046 (-2.392)**	0.00281 (1.677)	0.50844 (0.898)	0.2387 1265
231x0 - G	-0.12710 (-2.828)*	0.00213 (2.935)*	-0.00333 (-0.657)	2.55459 (3.501)*	0.8761 70
272x0 - G	0.05189 (6.741)*	-0.00068 (-5.228)*	0.00175 (2.018)**	0.32425 (2.185)**	0.3320 719
542x0 - E	0.05852 (3.387)*	-0.00094 (2.868)*	0.00257 (1.051)	0.39874 (1.665)***	0.3798 428
472x2 - M	0.03120 (1.636)	-0.00033 (-0.896)	0.01451 (4.383)*	-0.54063 (-2.077)*	0.4638 309
316x3 - E	0.13832 (1.674)	-0.00129 (-0.903)	0.00111 (0.072)	-1.36001 (-0.522)	0.5397 49
122x0 - G	0.06014 (4.947)*	-0.00090 (-4.254)*	0.00382 (2.762)*	0.20542 (0.099)	0.4346 676
732x0 - A	0.07000 (3.830)*	-0.00072 (-2.192)**	0.00883 (3.481)*	0.43517 (-0.160)	0.2011 1730
201x0 - G	0.08072 (3.203)*	-0.00134 (-2.803)*	-0.00277 (-0.695)	0.45273 (1.206)	0.0588 200

Table 22. Continued

AFS - AI	TAFMS	TAFMS ²	Aptitude	Constant	R-square/obs
702x0 - A	0.03745 (2.453)**	-0.00027 (-1.005)	0.00256 (1.179)	0.49007 (0.094)	0.2094 894
651x0 - G	0.04515 (0.640)	-0.00010 (-0.076)	0.01944 (2.083)**	-0.60465 (-0.767)	0.5352 145

* - $p < .01$

** - $p < .05$

*** - $p < .10$

CONCLUSIONS

This research effort investigated the potential for developing a new measure of enlisted airman productivity based upon Air Force occupational survey data. Occupational survey data were chosen because of the low cost associated with test development and data collection relative to the JPM data collection program. The new measure of productivity was developed from the OMS data based upon the concept of a set of core tasks existing for each AFS which can be used as a benchmark for determining the productivity of enlisted personnel.

Past productivity/performance measures were derived from surveys which had few career fields and a limited number of task performance data collected over an extended period of time with a large staff of raters and evaluators. The use of the OMS data as the source for the development of the productivity measure (OPI) potentially allows productivity functions to be estimated for each 5-digit AFS owing to the availability of OMS data for all AFSs. Large numbers of AFSs are surveyed annually in the OMS program allowing productivity functions to be frequently updated for the AFSs. The OMS data also contain respondent data for all tasks within an AFS (as opposed to the limited sample of tasks selected for the JPM study). The inclusion of all tasks within an AFS allows performance measures to be calculated at any level of aggregation for an AFS, e.g., by skill level or job.

The OPI productivity/proficiency measure developed in this effort was validated through its ability to rank order individuals similar to the more well-established measure of productivity in the JPM data. The rank order statistics showed that the OPI measure rank ordered individuals similar to the JPM measure. Aptitude and experience were statistically significant predictors of OPI. This finding was consistent with the findings of Nathan (1992). The E-Value method for determining core tasks was also found to be an excellent surrogate for the SME method, providing a lower cost method of determining core tasks for each AFS. In all, the OPI measure showed promising results as a job performance measure for enlisted personnel.

Table 23. Standardized Coefficient Values for OPI

AFS - AI	Experience	Aptitude
461x0 - M	0.0082	0.0046
492x1 - A	0.0351	-----
426x2 - M	0.0149	0.0065
324x0 - E	0.0433	0.0081
811x0 - G	0.0076	0.0058
902x0 - G	0.0111	0.0035
603x0 - M	0.0218	-----
231x0 - G	0.2709	-----
272x0 - G	0.0386	0.0047
542x0 - E	0.0188	-----
472x2 - M	-----	0.0208
316x3 - E	-----	-----
122x0 - G	0.0118	0.0048
732x0 - A	0.0189	0.0052
201x0 - G	0.0323	-----
702x0 - A	0.0369 ⁺	-----
651x0 - G	-----	0.0131

⁺ - Includes coefficient for TAFMS only.

RECOMMENDATIONS

The theory supporting the OPI measure could be further validated through surveys given to SMEs inquiring about the methodology used in allocation of tasks among their work force and their perception of task allocation as it related to the experience spectrum of their workers. The construct of the OPI measure should be further researched with respect to the weighting of each of the four factors which comprise the measure. Present equal weighting may not be the best alternative for capturing the relationships expressed in the productivity function.

The OPI measure should be calculated for each 5-digit AFS for which OMS data are available in addition to the 17 investigated in this effort. Productivity functions should be estimated for each AFS to determine the generalizability of the OPI performance measurement methodology to other AFSs. Further research should also be directed towards the functional form of the relationship between job performance and aptitude and experience. Efforts should also be made to extend the OPI methodology to airmen beyond the first-term. Further research is also warranted concerning the generalizability of the OPI measure to AFSs comprised of heterogeneous task structures. Additional research concerning the application of OPI at the job level within AFS could also provide better measures of job performance.

The E-Value methodology should be further researched to determine if other core task identification methodologies can be developed which would do a better job of identifying core tasks. The core tasks identified by the E-Value method should be validated for at least a subsample of AFSs through the use of SME's to determine if the tasks identified are truly tasks of a core nature to the AFS.

REFERENCES

- Army Research Institute (ARI), Human Resources Research Organization (HumRRO), Personnel Decisions Research Institute (PRDI), American Institutes for Research (AIR) (1986). *Improving the selection, classification, and utilization of army enlisted personnel: annual report* (ARI-TR-813101). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.
- Arnold, H.J. (1982). Moderator variables: a clarification of conceptual, analytic, and psychometric issues. *Organizational Behavior and Human Performance*, 34, 214-224.
- Arnold, H.J. (1984). Testing moderator variable hypotheses: a reply to Stone and Hollenbeck. *Organizational Behavior and Human Performance*, 42, 246-249.
- Arrow, K.J., Chenery, H.B., Minhas, B.S., & Solow, R.M. (1961). Capital labor substitution and economic efficiency. *The Review of Economics and Statistics*, August 1961, 225-250.
- Alley, W.E., & Teachout, M.S. (1990, August). *Aptitude and experience trade-offs on job performance*. Paper presented at the American Psychological Association Convention, Boston, MA.
- Barrett, R.S. (1966). *Performance Rating*. Science Research Associates, Inc.
- Bass, M., Bernard, M., & Barrett, G.V. (1981). *People, work, and organizations: an introduction to industrial and organizational psychology*. Boston: Allyn and Bacon, Inc.
- Becker, G.S. (1971). *Economic theory*. New York: Alfred A. Knopf.
- Becker, G.S. (1965). A theory of the allocation of time. *Economic Journal*, September 1965.
- Becker, G.S. (1964). *Human capital*. New York: National Bureau of Economic Research; and New York: Columbia University Press.
- Benjamin, R., Jr (1952). A survey of 130 merit-rating plans. *Personnel*, 29, 289-294.
- Carpenter, M.A., Monaco, S.J., O'Mara, F.E., & Teachout, M.S. (1989). *Time to job proficiency: a preliminary investigation of the effects of aptitude and experience on productive capacity* (AFHRL-TP-88-17, AD-210 575). Brooks AFB, TX: Training systems Division, Air Force Human Resources Laboratory.
- Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*. 52, 81-302.

- Eichel, E., & Bender, H.R. (1984). *Performance appraisal: a study of current technique*. American Marketing Association Research and Information Service.
- Faneuff, R.S., Valentine, L., Stone, B.M., Curry, G.L., & Hageman, D.C. (1990). *Extending the time to proficiency model for simultaneous application to multiple jobs* (AFHRL-TP-90-42). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.
- Flamholtz, E. (1985). *Human resource accounting*. San Francisco, CA: Jossey-Bass.
- Ford, J.K., Sego, D., & Teachout, M.S. (1991, April). *A test of the influence of general ability, job experience, and task experience on task level performance*. Paper presented at the Sixth Annual Conference of the Society for Industrial and Organizational Psychology, St. Louis, MO.
- Guion, R.M. (1965). *Personnel testing*. New York: McGraw-Hill.
- Harris, D.A., McCloy, R.A., Dempsey, J.R., Roth, C., Sackett, P.R., Hedges, L.V., Smith, D.A., & Hogan, P.F. (1991). *Determining the relationship between recruit characteristics and job performance: a methodology and a model* (HumRRO FR-PRD-90-17). For the Office of the Secretary of Defense, Force Management and Personnel. Alexandria, VA: Human Resources Research Organization.
- Hedge, J.W., & Teachout, M.S. (1986). *Job performance measurement: a systematic program of research and development* (AFHRL-TP-86-37, AD-A175 175). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.
- Heneman, H.G. III, & Schwab, D.P. (1972). Evaluation of research on expectancy theory predictions of employee performance. *Psychological Bulletin*, 78, 1-9.
- Hicks, J.R. (1932). Hicksonian neutrality in technological change. P.H. Douglas, *Theory of Wages*. London: Macmillan.
- Horowitz, S.A., & Sherman, A. (1980). A direct measure of the relationship between human capital and productivity. *Journal of Human Resources*, 15, 67-76.
- Intriligator, M.D. (1978). *Econometric models, techniques, & applications*. Englewood Cliffs, NJ: Prentice-Hall.
- Kirchner, W., & Reisberg, D.J. (1962). Differences between better and less effective supervisors in appraisal of subordinates. *Personnel Psychology*, 15, 295-302.
- Korman, A.K. (1970). *Industrial and organization psychology*. Englewood Cliffs, NJ: Prentice-Hall.

- Lance, C.E., Hedge, J.W., & Alley, W.E. (1987). *Ability, experience, and task difficulty predictors of task performance* (AFHRL-TP-87-14). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.
- Landy, F. J., & Trumbo, D.A. (1980). *Psychology of work behavior*. Homewood, IL: The Dorsey Press.
- Lipscomb, M.S., & Hedge, J.W. (1988). *Job performance measurement: Topics in the performance measurement of enlisted personnel* (AFHRL-TP-87-58, AD-A195 630). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.
- Mendenhall, W. & Scheaffer, R.L. (1973). *Mathematical statistics with applications*. North Scituate, MA: Duxbury Press.
- Miller, M.R., Skinner, J., & Harville, D.L. (1992). Utility of occupational surveys for assessing enlisted job performance. *Proceedings of the 34th Annual Conference of the Military Testing Association*. 2, 821-826.
- Nathan, B.R. (1992, May). *Job progression and position appraisal: Two relevant and available criteria for test validation*. Paper presented at the Seventh Annual Meeting of the Society of Industrial and Organizational Psychology, Montreal, Quebec, Canada.
- Nathan, B.R., & Nathan, M.L. (1991, April). *Number of tasks performed as a criterion for test validation*. Paper presented at the Sixth Annual Meeting of the Society of Industrial and Organizational Psychology, St. Louis, MO.
- Office of the Secretary of Defense (OASD; 1982-1989, annual); Manpower, Installations, and Logistics. *Joint service efforts to link enlistment standards to job performance, annual report to the house committee on appropriations*. Washington DC: Office of the Secretary of Defense.
- Phalen, W.J., & Weissmuller, J.J. (1981). CODAP: Some new techniques to improve job type identification and definition. *Proceedings of the 23rd Annual Conference of the Military Testing Association*. 2, 939-955.
- Porter, L.W., & Lawler, E.E. (1968). *Managerial attitudes and performance*. Homewood, IL: Irwing-Dorsey.
- Saal, F.E., & Landy, F.J. (1977). The mixed standard rating scale: an evaluation. *Organizational Behavior and Human Performance*, 18, 19-35.
- Schmidt, F.L., Hunter, J.E., & Outerbridge, A.N. (1986). Impact of job experience and ability on job knowledge work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology*, 71, 432-439.

- Schmidt, F.L., Hunter, J.E., Outerbridge, A.N., & Goff, S. (1988). Joint relation of experience and ability with job performance: test of three hypotheses. *Journal of Applied Psychology*, 73, 46-57.
- Stacey, W.D., Weismuller, J.J., Barton, B.B., & Rogers, C.R. (1974). *CODAP: Control card specifications for the UNIC 1108* (AFHRL-TR-74-84). Brooks AFB, TX: Computational Sciences Division, Lackland AFB, TX.
- Stone, B.M. (1989). *Time to proficiency model to link job performance and enlistment standards: Calculation of productive capacity and establishment of relations between productive capacity, ASVAB composites, and experience levels*. Unpublished manuscript, RRC, Inc.: Bryan, TX.
- Stone, B.M., Rettenmaier, A.J., Saving, T.R., & Looper, L.T. (1989). *Cost-based value models of Air Force experience* (AFHRL-TP-89-20, A212 771). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.
- Stone, B.M., Turner, K.L., Engquist, S.K., & Looper, L.T. (1992). *Simulation utility management system (SUMS): user's manual* (AL-TP-1992-0028). Brooks AFB, TX: Human Resources Directorate, Manpower and Personnel Division, Armstrong Laboratory.
- Stone, B.M., Turner, K.L., Fast, J.C., Curry, G.L., Looper, L.T., & Engquist, S.K. (1991). *A computer simulation modeling approach to estimating utility in several Air Force specialties* (AL-TR-1992-0006). Brooks AFB, TX: Human Resources Directorate, Manpower and Personnel Division, Armstrong Laboratory.
- Thorndike, E.L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, 4, 25-29.
- Vance, R.J., MacCallum, R.C., Coover, M.D., & Hedge, J.W. (1988). Construct validity of multiple job performance measures using confirmatory factor analysis. *Journal of Applied Psychology*, 73, 74-80.
- Vance, R.J., MacCallum, R.C., Coover, M.D., & Hedge, J.W. (1989). Construct models of task performance. *Journal of Applied Psychology*, 43, 447-455.
- Vroom, V.H. (1964) *Work and motivation*. New York: Wiley Press.
- Wexley, K.N. & Yukl, G.A. (1977). *Organizational behavior and personnel psychology*. Homewood, IL: Richard D. Irwin.
- Widaman, K.F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement*, 9, 1-26.

- Wiggins, V.L., Looper, L.T., & Engquist, S.K. (1991). *Applying neural networks to Air Force personnel analysis* (AL-TR-1991-0118). Brooks AFB, TX: Human Resources Directorate, Manpower and Personnel Division, Armstrong Laboratory.
- Wiley, L.N. (1976). *Airman job performance estimated from task performance ratings* (AFHRL-TP-76-64). Brooks AFB, TX: Occupation and Manpower Research Division, Air Force Human Resources Laboratory.
- Wiley, L.N. & Hahn, C.P. (1977). *Task level job performance criteria development* (AFHRL-TR-77-75). Brooks AFB, TX: Occupation and Manpower Research Division, Air Force Human Resources Laboratory.